On Knowledge Transfer in Object Class Recognition

A dissertation approved by
TECHNISCHE UNIVERSITÄT DARMSTADT
Fachbereich Informatik

for the degree of
Doktor-Ingenieur (Dr.-Ing.)

presented by

MICHAEL STARK
Dipl.-Inform.

born in Mainz, Germany

Prof. Dr.-Ing. Michael Goesele, examiner
Prof. Martial Hebert, Ph.D., co-examiner
Prof. Dr. Bernt Schiele, co-examiner

Date of Submission: 12th of August, 2010
Date of Defense: 23rd of September, 2010

Darmstadt, 2010
D17

ABSTRACT

In recent years, impressive results have been reported for the recognition of individual object classes, based on the combination of robust visual features with powerful statistical learning techniques. As a result, the simultaneous recognition of many object classes is coming into focus, posing challenges with respect to both model complexity and the need for increasing amounts of training data. Reusing once acquired information in the context of related recognition tasks, effectively transferring knowledge between object classes, has been identified as a promising route towards scalable recognition. Besides increasing scalability, knowledge transfer has been shown to enable novel tasks, such as the recognition of object classes for which no training data are available, termed zero-shot recognition. In this case, missing training data is compensated by exploiting additional, complementary sources of knowledge, such as linguistic knowledge bases. Based on these encouraging prospects, this thesis explores four different dimensions of knowledge transfer in object class recognition.

First, we investigate the role of visual features as a low level representation of transferable knowledge. Based on an extensive evaluation of existing state-of-the-art local feature detectors and descriptors, we identify shape-based features in connection with powerful spatial models as a promising candidate representation. Building upon this result, we further introduce a novel flavor of local shape-based features, as well as a generic appearance descriptor based on shading artifacts.

Second, we highlight the connection between knowledge transfer and generalization across basic-level object categories, by recognizing objects according to potential functions or affordances. In particular, we demonstrate that visually distinct hints on affordances, modeled as collections of local shape features, can be shared and hence transfered between object classes.

Third, we design shape-based object class models for knowledge transfer, representing object classes as spatially constrained assemblies of parts, including pair-wise symmetry relations. These models are both compositional and incremental, allowing for knowledge transfer either on the level of entire object class models or restricted to a subset of model components. While knowledge transfer in these models has to be guided by manual supervision, we demonstrate the benefit of knowledge transfer for object class recognition when learning from scarce training data.

And fourth, we demonstrate that exploiting additional sources of knowledge besides real world training images can aid object class recognition, effectively transferring knowledge between different representations. In particular, we use linguistic knowledge bases in connection with semantic relatedness measures to automatically determine potential sources and targets of knowledge transfer for zero-shot recognition, and show the successful learning of shape-based object class models from collections of 3D computer aided design (CAD) models, not using any real world

training images of the object class of interest.

In summary, this thesis achieves encouraging results with respect to four different dimensions of knowledge transfer, namely, specialized visual feature representations, generalization across basic-level categories, compositional object class models, and the exploitation of additional sources of knowledge, confirming the benefits of knowledge transfer. As a side effect, we are able to obtain object class recognition results often superior to or en par with prior work.

ZUSAMMENFASSUNG

In den letzten Jahren wurden bemerkenswerte Ergebnisse im Erkennen einzelner Objektklassen erzielt, erreicht durch die Kombination von robusten visuellen Merkmalen mit Verfahren des statistischen maschinellen Lernens. In der Folge rückt das simultane Erkennen vieler Objektklassen in den Fokus, was Herausforderungen sowohl hinsichtlich der Modellkomplexität als auch der Menge der benötigten Trainingsdaten mit sich bringt. Wiederverwendung und Transfer von einmal gewonnenem Wissen zwischen verwandten Erkennungsaufgaben wurde als ein vielversprechender Ansatz zum Erreichen skalierbarer Erkennung erkannt. Dabei ermöglicht Wissenstransfer neben gesteigerter Skalierbarkeit das Lösen neuartiger Aufgaben, wie etwa das Erkennen von Objektklassen, für welche keine Trainingsdaten verfügbar sind, gennant zero-shot recognition. In diesem Falle werden fehlende Trainingsdaten durch das Heranziehen zusätzlicher, komplementärer Wissensquellen ersetzt, zum Beispiel linguistischer Natur. Inspiriert vom Potenzial des Wissenstransfers untersucht diese Arbeit vier verschiedene Richtungen des Wissenstransfers im Erkennen von Objektklassen.

Die erste Richtung untersucht die Rolle von visuellen Merkmalen als die Repräsentation von transferierbarem Wissen auf der untersten Abstraktionsebene. Als Basis dient eine umfangreiche Evaluation verschiedener lokaler Merkmalsextraktoren und -Deskriptoren, welche formbasierte Repräsentationen in Kombination mit ausdrucksstarken räumlichen Modellen als vielversprechend identifiziert. Diesem Resultat folgend entwickeln wir weiters eine neuartige Variante einer formbasierten Repräsentation und einen generischen Deskriptor zur Charakterisierung von Oberflächenschattierungen.

Die zweite Richtung beleuchtet die Verbindung zwischen Wissenstransfer und der Generalisierung zwischen Kategorien der Basisebene (basic-level categories), am Beispiel des Erkennens funktionaler Objektklassen. Insbesondere verdeutlichen wir, dass unterschiedlichen Objektklassen visuelle, formbasierte Merkmale gemein sein können, welche auf potenzielle Funktionen (sogenannte affordances) hinweisen. Jene Merkmale sind folglich zwischen den Objektklassen transferierbar.

Die dritte Richtung ist dem Entwurf formbasierter Objektklassenmodelle gewidmet, welche Objektklassen als Ansammlungen von Teilen in einer festgelegten räumlichen Anordnung beschreiben, und zusätzlich paarweise Symmetriebeziehungen zwischen Paaren von Teilen einbeziehen. Jene Modelle sind gleichzeitig komponierbar und inkrementell erweiterbar, und erlauben somit Wissenstransfer auf der Ebene vollständiger Modelle und auf der Ebene von Teilmodellen. Obwohl der Wissenstransfer in diesen Modellen von Hand spezifiziert werden muss, zeigt sich der Nutzen des Wissenstransfers im Falle weniger verfügbarer Trainingsdaten.

Die vierte Richtung demonstriert die Verwendung von zusätzlichen Wissensquellen zur Verbesserung der Objektklassenerkennung, indem Wissen zwischen unter-

schiedlichen Repräsentationen transferiert wird. Insbesondere untersuchen wir die Verwendung linguistischer Wissensquellen in Verbindung mit Maßen der semantischen Verwandtschaft, um automatisch potenzielle Wissenstransferquellen und -Ziele zu bestimmen. Weiters zeigen wir das erfolgreiche Lernen formbasierter Objektklassenmodelle aus einer Sammlung von 3D computer aided design (CAD-) Modellen, wobei wir auf jegliche Trainingsbilder der jeweiligen Objektklasse verzichten.

Insgesamt erzielt diese Arbeit vielversprechende Resultate bezüglich vier verschiedener Richtungen des Wissenstransfers: spezialisierte Repräsentationen visueller Merkmale, Generalisierung zwischen Kategorien der Basisebene, komponierbare Objektklassenmodelle, und die Verwendung zusätzlicher Wissensquellen. Als Nebeneffekt wird eine oft bessere oder gleichwertige Performanz verglichen mit früheren Arbeiten in der Objektklassenerkennung erzielt.

ACKNOWLEDGEMENTS

First and foremost, I want to thank Prof. Bernt Schiele for supervising my thesis, and being a constant source of inspiration and motivation throughout the time. In particular, I am grateful for his confidence in my abilities from the beginning, which allowed me to grow from a layman in computer vision to being fluent in object class recognition. Likewise, I thank Prof. Michael Goesele for co-supervising my thesis and agreeing to serve as an examiner. I am more than thankful for his advice, which often proved invaluable, precisely because of its non-vision perspective. I am truly grateful to Prof. Martial Hebert for serving as an external reviewer as part of the thesis committee.

I would also like to express my gratitude to all members of the MIS, IU, ESS, and GRIS groups, not only for supporting me with inspiring discussions and feedback concerning research, but also sharing a lot of fun moments in leisure activities: Jens Ackermann, Anton Andriyenko, Eugen Berlin, Ulf Blanke, Marko Borazio, Victoria Carlsson, Dr. Gyuri Dorko, Simon Fuhrmann, Dr. Tam Huynh, Nikodem Majer, Kevin Schelten, Paul Schnitzspan, Dr. Edgar Seemann, Dr. Ulrich Steinhoff, Christoph Vogel, Stefan Walk, Dr. Maja Stikic, and Dr. Andreas Zinnen. I owe particular thanks to Ursula Paeckel, for being the good soul of the group, and having a sympathetic ear for all matters, and my office mates Dr. Christian Wojek and Micha Andriluka for many fruitful discussions and out of line thinking. I thank Dr. Mario Fritz and Dr. Kristof van Laerhoven for sharing both their expertise and sense of humor, and Dr. Diane Larlus for saving my life in Kyoto with her Japanese.

Furthermore, I would like to thank my collaborators, without whom I would not have had the chance to complete this thesis: Prof. Iryna Gurevych, Philipp Lies, Marcus Rohrbach, Prof. Konrad Schindler, Dr. György Szarvas, Dr. Jeremy Wyatt, Dr. Michael Zillich, and Zeeshan Zia. Similarly, my thanks go to the students that I had the opportunity to supervise and work with, Sebastian Schneider and Sandra Ebert.

I further thank the EU project CoSy and the DFG for providing both funding and an exciting context for my research, and allowing me to collaborate with many great researchers around the globe.

Lastly, I am very grateful to many people that encouraged me to continue my way in research, especially Dr. Patrick Lehti, Dr. Peter Fankhauser, Dr. Mary F. Fernández, and Dr. Jérôme Siméon. Most of all, I thank my family, in particular my parents, and my dear friends Gerald Bork and Wolfgang Lennartz for always believing in me, and pulling me back onto the ground in testing times.

CONTENTS

1 INTRODUCTION

Contents

UNDERSTANDING visual scenes is one of the most important and remarkable abilities of the human cognitive system. Humans can process and interpret complex visual scenes at the blink of an eye, including the recognition of other individuals, objects, and their interactions (Gibson, 1979). Despite the rapid development of computing technology, machine vision is still far from the versatility of the human vision system. As a consequence, related research has typically focused on sub-tasks or simplified variants of the scene understanding problem. In recent years, tremendous progress has been made in the field of object class recognition. Based on robust local features in connection with powerful machine learning techniques, remarkable recognition performance has been demonstrated on a wide variety of object classes, as, for instance, in the course of the PASCAL visual object classes (VOC) challenges (Everingham *et al.*, 2010). While early approaches to object class recognition had been tied to simplistic settings concerning both recognizable objects and scenes (Nevatia and Binford, 1977; Marr and Nishihara, 1978; Brooks *et al.*, 1979; Pentland, 1986; Lowe, 1987), modern recognition systems successfully cope with many of the challenges posed by real world imagery, such as intra class variation, cluttered backgrounds, lighting variations, and even partial occlusion (Fergus *et al.*, 2003; Leibe *et al.*, 2006a; Felzenszwalb *et al.*, 2009). These achievements have been made due to both advances in the robust encoding of real world image information and the development of learning algorithms capable of separating relevant object class information from background noise, given sufficient representative training data of both the object class of interest and unrelated background.

Having reached good performance on the level of individual object classes, the even more challenging task of recognizing many object classes simultaneously is getting into focus (e.g., in the ImageNet Large Scale Visual Recognition Challenge [1] (Everingham *et al.*, 2010)). Besides increasing the computational cost, recognizing many

[1] http://www.image-net.org/challenges/LSVRC/2010/

classes poses the additional challenge of acquiring enough training data, which in many cases have to be accompanied by annotations of varying granularity in order to be usable by recognition algorithms, such as bounding boxes hinting the position of objects or even pixel by pixel labelings. As a consequence, numerous approaches have been proposed that explicitly aim at reducing required annotations based on semi-supervised or completely unsupervised machine learning techniques (Weber *et al.*, 2000; Fergus *et al.*, 2003). While these approaches effectively limit the amount of required annotations for the case of single object class recognition, scaling these approaches to higher numbers of classes has proven difficult due to their often limited discriminative power in comparison to supervised methods.

As a consequence, making efficient use of once acquired information has been recognized as a promising route towards scaling recognition to higher numbers of classes. In particular, sharing information between object classes on the level of individual features (Torralba *et al.*, 2004) or entire object class models (Fink, 2004; Bart and Ullman, 2005b), effectively *transferring knowledge* from one to another, is considered key to success. Besides the need for an appropriate representation of transferable knowledge, knowledge transfer requires an understanding of the structure underlying the space of object classes (Zweig and Weinshall, 2007; Marszalek and Schmid, 2007), in order to determine possible sources and targets of transfer. Therefore, this thesis investigates both suitable object class representations and mechanisms for the automatic determination of sources and targets of knowledge transfer, in the context of object class recognition.

1.1 KNOWLEDGE TRANSFER IN OBJECT CLASS RECOGNITION

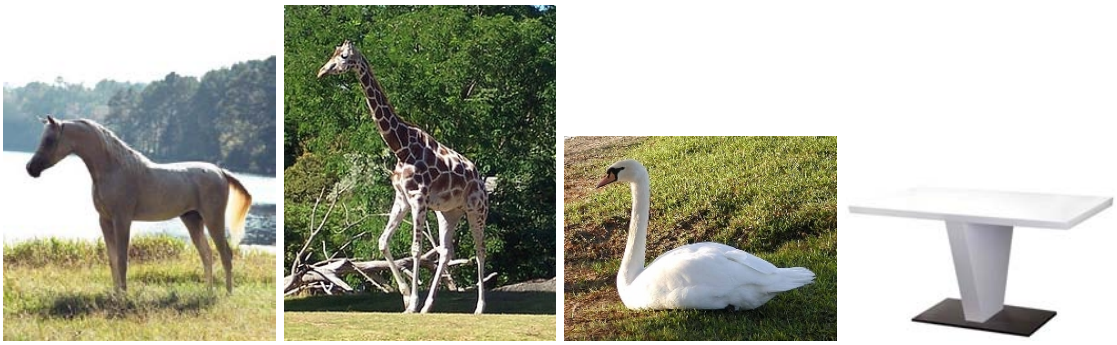


Figure 1.1: Knowledge transfer in object class recognition.

The central question that will be pursued in this thesis is visualized in Figure 1.1. Assuming that an object class model for the class *horse* is available, in what way can it facilitate the recognition of other object classes, such as *giraffe*, *swan*, or *dining table*? Intuitively, the existing model should be particularly helpful for the class giraffe, since both horses and giraffes are quadrupeds, sharing visually similar

statures, despite exhibiting different proportions, and being differently textured. In particular, both share a common set of similarly shaped body parts, arranged in a common layout. Since the swan is an animal, it shares at least a subset of body parts with the horse, such as torso, neck, head, etc. Again, the proportions deviate largely from those of the horse, and its feathering generates a different appearance than the horse's hide. Surprisingly, although probably least of all, even the recognition of dining tables might profit from an existing horse model. While both appearance and shape are far from any of the animal classes, it still shares the quadrupeds' functional relation between a resting body and a supporting leg. While this functional relation does not have any direct visual manifestation in terms of appearance or shape, it is reflected in the geometric arrangement of body, legs, and ground.

For the animal example, it seems intuitive that knowledge transfer on the level of distinct visual properties, such as texture, color, or even entire object parts, is potentially useful. This intuition has in fact inspired the creation of an animal image data set (Animals-with-Attributes, AwA (Lampert *et al.*, 2009)), designed particularly for the challenging task of zero-shot recognition, i.e., recognizing animals for which an abstract description in terms of visual properties (attributes) exists, but zero training images. Knowledge transfer comes into play when recognizers for these visual properties are trained from one set of animal images, but tested on another. Taking the idea of attribute-based recognition further, others advocate a paradigm change for categorization (into object classes) to description (by attributes) (Farhadi *et al.*, 2009), implicitly transferring knowledge between object classes by re-creating them from more generic visual building blocks. Besides the possibility for zero-shot recognition, object class descriptions offer the advantage of increased representational power, allowing to specify relevant properties in addition to category affiliation (e.g., "the red car", "a yellow dress" (Wang and Forsyth, 2009), or "a face with a similar nose than Harrison Ford's" (Kumar *et al.*, 2009)).

Even when retaining the well-established category-based recognition paradigm, knowledge transfer offers the potential of significantly reducing both model complexity and the number of needed training examples, as has been demonstrated in the context of the joint learning of multiple object classes (Ben-David and Schuller, 2003; Torralba *et al.*, 2004). Following this direction, distance-based approaches have achieved remarkable recognition performance by using as few as a single training example per object class (Fink, 2004; Bart and Ullman, 2005b). Re-using knowledge from related classes apparently constitutes an effective measure against overfitting (Thrun, 1996), thus providing increased generalization ability. Similar observations have been made when combining classifiers from different levels of a given object class hierarchy (Zweig and Weinshall, 2007; Marszalek and Schmid, 2007).

In addition to purely visual information, knowledge transfer also comprises the exploitation of additional knowledge sources, and their connection with visual information. In the literature, attempts have been made mostly to combine visual with textual information, such as image captions or tags. The resulting joint models have been demonstrated to yield superior performance compared to using image

information alone, and may pave the way to a more holistic understanding of visual scenes (Li *et al.*, 2009; Jamieson *et al.*, 2010). Reviving ideas from the early days of computer vision, 3D models from computer graphics applications have recently been suggested as another useful source of information (Liebelt *et al.*, 2008; Liebelt and Schmid, 2010). In contrast to real world image training data, 3D models have the advantage of providing both accurate object boundaries and geometry, optionally including textured surfaces, and being inherently viewpoint invariant.

Conceptually, transferring knowledge between object classes is similar to the notion of generalization beyond basic-level categories (Rosch *et al.*, 1976), which has been around for some time (Winston *et al.*, 1983; Stark and Bowyer, 1991), and since has inspired mostly works in robotics-related literature (Kjellström *et al.*, 2008).

To conclude, knowledge transfer offers vast potential to aid object class recognition, both in terms of enabling new tasks, such as zero-shot recognition, and reducing the computational complexity of tasks involving many object classes. In this thesis, we will substantiate this potential by contributing to different aspects of knowledge transfer in object class recognition, including specialized image feature representations (chapters 3 and 6), compositional object class models (chapters 4 and 5), and exploitation of additional knowledge sources, specifically 3D computer aided design (CAD) models (Chapter 7) and linguistic knowledge bases (Chapter 8).

1.2 CHALLENGES FOR KNOWLEDGE TRANSFER

While the animal example of Section 1.1 illustrates some of the potentials of knowledge transfer in object class recognition, it also hints on several of its challenges. In this section, we highlight and discuss these challenges. We first give an overview of the challenges that apply to object class recognition in real world images in general, and then move on to the challenges specific to knowledge transfer.

1.2.1 Challenges for object class recognition in general

Object class recognition is mostly challenged by the discrepancy between the depiction of an object in a real world image and its corresponding abstraction in an object class model. The following challenges are well known from the literature, and have been tackled by prior work to a varying degree. The work presented in this thesis is in principle susceptible to all of them. Section 1.3.1 highlights the contributions of this thesis with respect to meeting the various challenges.

Intra-class variation vs. inter-class discrimination. One of the main challenges of object class recognition is the tradeoff between intra-class variation and inter-class discrimination. On the one hand, object class models have to be general enough to capture the variability between different instances of an object class, which potentially ranges from small deviations in local texture to substantial differences in global geometry. On the other hand, an object class model must

be specific enough to precisely capture those aspects that discriminate the object class of interest from background or other object classes.

Cluttered backgrounds. Foreground objects can be hard to detect even to the human eye, if they are located in front of a highly structured (cluttered) background. The challenge consists in separating these accidental background structures from non-accidental structures characteristic of the object class of interest.

Varying viewpoints. Even when looking at a single instance of a given object class, there is an infinite number of possible different depictions of this instance due to varying viewpoint. The resulting change in appearance can be substantial.

Varying lighting conditions. A similar variability in appearance originates from varying lighting conditions. The challenge consists in designing object class representations that are invariant to these variations, by abstracting from the original photometric measurements, and focusing on relevant appearance or shape information instead.

Articulation and pose variation. An additional source of appearance variation originates from the non-rigid nature of many object classes, such as humans or animals. Instances of non-rigid object classes allow for certain deformations, and, consequently, different poses in which instances can be depicted.Object class models are thus required to exhibit flexibility in order to cope with these deformations.

Partial occlusion. Lastly, foreground objects may be visible only partially due to occlusion. The occlusion can either be caused by parts of the object itself (self occlusion) or other objects. In both cases, object class recognition must infer the presence of the object class of interest from reduced evidence, rendering discrimination from background more difficult.

1.2.2 Challenges specific to knowledge transfer

Integrating knowledge transfer into object class recognition brings along its own challenges, mostly related to the representation of transferable knowledge, and the determination of how it should be used. Section 1.3.2 highlights the contributions of this thesis with respect to meeting the challenges specific to knowledge transfer.

Identifying transferable knowledge. The most basic requirement for knowledge transfer between object classes is naturally the identification of transferable properties. Ideally, these properties should fulfill three requirements. First, each property should be visually distinctive (otherwise, it is simply not accessible by visual processing). Second, properties should be shared among several object classes in order to be practically transferable. And third, each property should be shared by a non-trivial subset of the classes, and the subsets induced

by different properties should be sufficiently diverse to provide a meaningful characterization of object classes.

The challenge consists in identifying transferable visual properties fulfilling the above requirements, and in particular, identifying these properties automatically, without the need for human supervision.

Representing transferable knowledge. Finding an appropriate representation of transferable knowledge, once identified on a conceptual level, poses an additional challenge. Some apparently transferable visual properties may not have obvious implementations in terms of visual feature representations, or may imply prohibitive computational complexity during recognition. As an example, let us come back to transferring knowledge between the object classes *horse* and *giraffe* (see Figure 1.1). While the similarity in the overall stature of the two quadruped classes is immediately apparent to human intuition, it is much harder to characterize that similarity in a formal model. One possible formalization is based on describing object classes as a spatially constrained assembly of parts, each of which is in turn characterized by its own distinct appearance. In this way, the common overall stature can be transfered separately, without being affected by the varying appearance of the individual parts.

Having defined a representation of transferable knowledge still leaves the question as to whether this representation can be handled efficiently in the context of object class recognition, leading to a tradeoff between richness in representation on one and efficient computability on the other hand.

Integrating transfer into object class models. Having found a suitable representation of transferable knowledge, this representation has to be integrated into an object class model in order to be useful. This in turn requires the object class model to support integration, either through compositionality, i.e., allowing to plug in additional components corresponding to transfered knowledge, or incrementality, i.e., allowing updates of its state in response to transfered knowledge. Furthermore, the ideal object class model facilitates the integration of various different kinds of transfered knowledge.

Determining sources and targets of transfer. Given a general means of representing transferable knowledge, it remains the question as to how to determine which knowledge to transfer where for a concrete set of object classes. While, for initial experimentation, manual supervision is acceptable, the goal is to automate the determination of possible sources and targets of knowledge transfer as much as possible. Technically, this poses the additional challenge of interfacing between a representation of transferable knowledge on the vision side with an abstract description that allows reasoning about sources and targets. This challenge is complicated by the potentially multiple different kinds of transfered knowledge supported by an ideal object class model.

Exploiting new knowledge sources. An alternative approach to transferring knowledge between object classes is tapping completely new sources of knowledge in order to learn object class models, besides using real world training images. Since the ultimate goal is the application of such an object class model in order to recognize real world objects, it remains the challenge of matching two different object representations. As an example, consider 3D computer aided design (CAD) models. While the accurate representation of three dimensional geometry clearly offers the potential of learning equally accurate spatial models, finding an appropriate representation for the appearance is non-trivial.

1.3 CONTRIBUTIONS OF THE THESIS

In this section, we highlight the contributions of the thesis, and relate them to each of the challenges discussed in Section 1.2. The thesis makes both contributions to the field of general object class recognition, and contributions specific to knowledge transfer. In particular, it demonstrates competitive recognition performance to prior work in a variety of standard benchmarks. Furthermore, it shows that knowledge transfer can effectively reduce the amount of required training data needed for learning object class models, as well as the amount of needed human supervision. Both of these achievements are important ingredients for scalable recognition.

Section 1.4 is also concerned with the contributions of this thesis, but in the form of a chronological walk-through of the individual chapters. The placement of the contributions with respect to prior work in the field is given in Chapter 2.

1.3.1 Contributions to object class recognition in general

Intra-class variation vs. inter-class discrimination. The challenging tradeoff between intra-class variation and inter-class discrimination is related to a variety of contributions of the thesis. First, we provide an extensive evaluation of various local shape and appearance features with respect to their ability to discriminate between different object classes (Chapter 3). The evaluation comprises the analysis of individual feature spaces as well as combinations with discriminative classifiers and spatial models. Second, we examine both representative and discriminatively trained models for the shape of object parts in a part-based object class model (chapters 5, 6, and 7). In both cases, we demonstrate competitive recognition performance compared to prior work on standard benchmark data sets.

Cluttered backgrounds. The object class models presented in chapters 4, 5, and 7 are entirely based on representing the shape of objects, either in the form of distinct edge segments (chapters 4, 5) or as localized histograms of weighted image gradients (Chapter 7). While both representations provide only weak discrimination between foreground objects and background on the level of individual features, we demonstrate that combining multiple features in a

part-based model in connection with powerful spatial constraints renders these representations highly robust to background clutter. Furthermore, we show that exploiting non-accidental Gestalt principles improves performance. In particular, we use symmetries as an additional valuable cue for recognition (Chapter 5).

Varying viewpoints. In Chapter 7, we devise a shape-based object class model that is particularly tailored towards recognizing object classes from multiple viewpoints, enabled by tapping an additional source of knowledge besides real world training images, namely, 3D computer aided design (CAD) models. Despite following the intuitive bank of detectors paradigm, we improve state-of-the-art multi-view recognition performance on a standard benchmark data set by a large margin.

Varying lighting conditions. Being based on shape rather than appearance, the object class models of chapters 4, 5, and 7 bypass the effects of varying lighting conditions in real world images to a fair degree. While this strategy is based on abstracting away unwanted effects, we follow a completely different route in Chapter 6. To this end, we contribute a novel representation of shading artifacts caused by interaction of light with curved surfaces, and use this representation as an additional cue for recognition. In fact, this additional cue can be beneficial on standard benchmark data, as we demonstrate in our experiments.

Articulation and pose variation. While the object class models presented in this theses have not been primarily designed for non-rigid object classes, they do provide robustness to pose variations to a varying degree. The part-based models of chapters 5 and 7 represent the spatial layout of object classes in a flexible, probabilistic fashion, assigning low but non-zero probabilities to previously unseen configurations. The local shape representations of these models can also account for small variations in part shape, resulting in competitive recognition performance even for the non-rigid object classes *giraffe* and *swan*.

The object class model of Chapter 8 is built along the lines of Lampert *et al.* (2009). As such, it is based on a collection of a multitude of visual features, subsumed in bag-of-visual-words histograms of varying spatial resolutions. As a consequence, this model is capable of representing local feature occurrences independent of their locations, resulting in state-of-the-art recognition performance on a benchmark data set of articulate animal object classes.

Partial occlusion. As for articulation and pose variation, partial occlusion is not a primary target of this thesis. Nonetheless, partial occlusion can be seen as an extreme case of weak local evidence, which is successfully handled by our object class models, as confirmed by our experiments. The part-based models of chapters 5, 6, and 7 are particularly amenable for handling weak local evidence, since it can be compensated for on a global level. While this thesis does not explore the handling of completely missing local evidence, we

note that reversible jump dynamics (Green, 1995) are a promising candidate solution for a clean implementation of partial occlusion handling as part of future work.

1.3.2 Contributions specific to knowledge transfer

Identifying transferable knowledge. While in chapters 4 and 5, transferable knowledge is identified by means of human supervision, Chapter 8 contributes a novel approach to automating this process. The approach is building on the existing idea of representing transferable knowledge in the form of binary classifiers. Each classifier divides the space of real world images into those images sharing a distinct visual property, and those that do not. Each classifier further draws from a large collection of different visual features, automatically selecting features relevant to its associated property during training. As a result, object classes can be represented as concurrences of multiple visual properties, corresponding to a combination of the respective classifier outputs. While prior work determines these combinations by manual supervision, the contribution of Chapter 8 consists in determining these combinations fully automatically, without the need for any human supervision.

As concerns the three requirements to visual properties listed in Section 1.2, the first one (visual distinctiveness) is met by definition, since properties are modeled as classifiers that select the most discriminative visual features during training (we note that, in theory, a classifier might pick up an incidentally correlated visual feature rather than one that truly corresponds to the property of interest (Farhadi *et al.*, 2009)). The second and third properties (balanced sharing of properties between object classes) are not enforced explicitly, but result from the specific means of generating an inventory of properties to choose from.

Representing transferable knowledge. In Chapter 5, we design a novel shape-based object class model specifically tailored towards knowledge transfer. It consists of a number of different components, each corresponding to a distinct visual property that can be transfered either individually or in combination with others. Object classes are represented as spatially constrained assemblies of parts, each being characterized by its shape and size relative to the assembly. In addition, pairwise relations between parts can be governed by symmetry descriptions. All components are probabilistic in nature, allowing for a clean integration into a joint density of an object class model.

As concerns computational complexity, Chapter 5 contributes an efficient approximate implementation of MAP search for recognition using the above described components, based on Markov Chain Monte Carlo sampling. This implementation is in principle applicable to arbitrarily complex densities, and guides the search to high density regions of the solution space by bottom-up proposals.

Integrating transfer into object class models. In a first line of research (Chapter 5), we propose an object class model that is both compositional and incremental. Compositionality is a consequence of the model corresponding to a joint probability density that factors into separate components, which can thus be exchanged, deleted, and appended. Compositionality can be applied to the shape of object parts, their relative sizes, their overall spatial layout, and the symmetry relations between pairs of parts. Incrementality is achieved by the specific functional form of the individual component densities, which allows to integrate prior information in the form of covariance estimates.

The object class model used in a second line of research (Chapter 8) is compositional in nature, since it represents object classes as combinations of individual classifier outputs, each corresponding to a distinct portion of transferable knowledge. The chosen formulation is again probabilistic, making use of factorization.

Determining sources and targets of transfer. In Chapter 8, we propose a novel approach to determining sources and targets of knowledge transfer fully automatically, without the need for any human supervision. Representing object classes as concurrences of visual properties, our approach determines the associations between these properties (sources of knowledge transfer) and object classes (targets of knowledge transfer) using their semantic relatedness. Semantic relatedness is computed on a variety of different linguistic knowledge bases, using semantic relatedness measures considered state-of-the-art by the natural language processing community. We further extend this work by additionally mining an inventory of visual properties automatically, again using semantic relatedness. The link between visual and language-based information is established through visual classifiers, trained from labeled training data (labels constitute language expressions).

In our experiments, we demonstrate that fully automatic knowledge transfer performs on par with human supervised knowledge transfer on a challenging benchmark data set for zero-shot recognition, which requires recognizing object classes for which zero real world training images are available.

Exploiting new knowledge sources. In Chapter 7, we propose to learn object class models solely from 3D CAD data, not using any training images of the object class of interest. This is in contrast to most prior work, which requires substantial amounts of real world training images either as the sole or additional source of information. In this respect, we provide a concrete measure of reducing the amount of needed image training data for scalable recognition. In particular, we propose an abstract shape representation that interfaces between 3D CAD and real world image data based on non-photorealistic rendering. We combine this representation with the powerful spatial model of Chapter 5, and demonstrate state-of-the-art recognition performance on a standard benchmark data set. Furthermore, we benefit from the viewpoint independence of 3D

CAD data, and extend the object class model to account for varying viewpoint, outperforming prior work by a large margin.

1.4 OUTLINE OF THE DOCUMENT

This section gives an overview of the organization of the thesis, following the chronological ordering of the constituent chapters. It provides a short summary of each chapter, together with information on the originating publications.

Chapter 2: Related work. This chapter gives an overview of related prior work in computer vision. It distinguishes among works in general object class recognition and works in the more specific field of knowledge transfer. It further categorizes related prior work according to employed visual feature representations, object class models, and particular flavors of knowledge transfer.

Chapter 3: Local features for classes of geometric objects. Visual feature representations constitute the basis of all object class recognition systems. They provide the abstraction necessary in order to make real world image data accessible through machine vision algorithms. The purpose of this chapter is to give an extensive evaluation of various state-of-the-art local feature detectors and descriptors, in a series of experiments related to object class recognition.

Since object shape is likely to be transferable between object classes, the evaluation is targeted towards shape representations. In particular, it compares recognition performance of various local shape and appearance features on both a standard benchmark data set for object class recognition (a subset of Caltech-101) and two newly proposed data sets (*Shape* and *Shape2*) featuring object classes that are characterized by shape rather than appearance. The results of the evaluation have inspired the design of the various shape-based object class models presented in later chapters.

The contents of this chapter corresponds to the ICCV 2007 publication "How Good are Local Features for Classes of Geometric Objects" (Stark and Schiele, 2007). In the context of cognitive systems, it has been published as part of the chapter "Categorical Perception" of the book "Cognitive Systems" (Fritz *et al.*, 2010).

Chapter 4: Functional object class detection. As pointed out earlier, knowledge transfer between object classes is closely related to the notion of generalization beyond categorical boundaries. While two objects may be instances of two different basic level categories, they might still belong to the same category on a more abstract level, e.g., related to object *affordances*. Affordances constitute properties of objects that afford to perform certain actions, and hence support certain functions. Having emerged as part of a robotics application in the context of EU project CoSy, the focus of this work lies on two variants of grasping actions.

Assuming that object affordances have visually perceivable counterparts in the form of groups of local shape features, we denote these features *affordance cues*. This chapter constitutes a first attempt at transferring knowledge between object classes, by demonstrating that affordance cues can be learned from objects of one basic level category, and rediscovered on objects of another.

The contents of this chapter corresponds to the ICVS 2008 publication "Functional Object Class Detection Based on Learned Affordance Cues" (Stark *et al.*, 2008), and has also been published as part of the chapter "Multimodal Learning" of the book "Cognitive Systems" (Skocaj *et al.*, 2010). The implementation is partly based on code developed in the course of the diploma thesis of Philipp Lies, "Extracting Affordance Cues from Observed Human-Object Interactions", which has been supervised by the author of this thesis. Parts of the code have further been integrated into the CoSy Architecture Schema Toolkit (CAST) framework (Hawes *et al.*, 2007).

Chapter 5: Shape-based object class model for knowledge transfer. While Chapter 4 has demonstrated the transferability of groups of local shape features, this chapter moves on to transferring diverse aspects of object geometry as well. It designs a shape-based object class model, which allows to transfer the shape of individual object parts, their spatial layout, and pairwise symmetry relations among them. The corresponding probabilistic formulation factors into separate components for each of these aspects, facilitating transfer of all aspects at once or just a subset, and even supporting transfer restricted to subsets of object parts. While the spatial layout component allows for the specification of a full covariance matrix governing part positions in the spirit of the *constellation model* (Fergus *et al.*, 2003), efficient approximate inference is achieved by adopting a Markov Chain Monte Carlo (MCMC) sampling scheme.

Since the focus of this chapter lies on the representational aspect of an object class model supporting knowledge transfer, it resorts to manual supervision in the specification of which knowledge should be transfered where. This limitation is addressed in Chapter 8 (in connection with a different object class model). The probabilistic formulation together with the proposed MCMC inference scheme provide the basis of the object class models described in chapters 6 and 7, respectively.

The contents of this chapter corresponds to the ICCV 2009 publication "A Shape-Based Object Class Model for Knowledge Transfer" (Stark *et al.*, 2009b).

Chapter 6: Shading cues for object class detection. While previous chapters have focused on shape as an example of transferable knowledge, this chapter introduces a novel kind of semi-local feature describing shading artifacts on object surfaces. Being based on a physical model of reflectance rather than on learning from examples, this feature is inherently generic in nature, and potentially transferable between object classes, although the experiments presented in this chapter are limited to a single object class of a standard recognition benchmark.

The proposed shading features are evaluated in combination with the shape-based object class model described in Chapter 5, and compared against a variant not using these features in terms of recognition performance.

The contents of this chapter corresponds to the 3dRR 2009 publication "Shading Cues for Object Class Detection" (Stark *et al.*, 2009a), which won a best paper award.

Chapter 7: Learning shape models from 3D CAD data. In contrast to earlier chapters, this chapter explores the use of alternative knowledge sources other than real world images for learning object class models. In particular, it revisits the idea of learning these models directly from 3D computer aided design (CAD) data, not using any training images of the object class of interest. In that sense, knowledge transfer is not taking place between object classes, but between different representations of one and the same object class.

The transition from 3D CAD data to real world images needed for recognition is realized by means of non-photorealistic rendering in connection with a robust local shape feature representation, and integrated into the probabilistic object class model of Chapter 5. Since 3D CAD data is inherently viewpoint independent, this model is instantiated multiple times for a set of discrete viewpoints. Recognition performance is consequently evaluated on a standard multi-view benchmark.

The contents of this chapter corresponds to the BMVC 2010 publication "Back to the Future: Learning Shape Models from 3D CAD Data" (Stark *et al.*, 2010).

Chapter 8: Semantic relatedness for knowledge transfer. A limitation of the approach presented in Chapter 5 is the required manual supervision that determines which knowledge to transfer where. This chapter thus investigates how possible sources and targets of knowledge transfer can be determined automatically, without human supervision. The proposed method uses publicly accessible knowledge sources, such as WordNet, Wikipedia, or web search engines, in connection with linguistic measures, to quantify the semantic relatedness between object classes. Knowledge is then transfered between classes according to their semantic relatedness.

The work presented in this chapter is based on two different representations of object classes. The first representation describes object classes as an assembly of visual attributes, implemented by means of attribute classifiers. The second representation characterizes classes by their semantic relatedness to a predefined set of reference classes. While both representations are clearly different from the object class model presented in Chapter 5 (which is particularly designed for transfer), they do allow for a direct comparison with prior work using the same representation (Lampert *et al.*, 2009).

The contents of this chapter corresponds to the CVPR 2010 publication "What Helps Where – And Why? Semantic Relatedness for Knowledge Transfer" (Rohrbach *et al.*, 2010), and has been approved as part of the diploma

thesis of Marcus Rohrbach, who has been co-supervised by the author of this thesis. The code as well as the experimentation related to this chapter have been provided by Marcus Rohrbach [2].

Chapter 9: Conclusions and future perspectives. In this chapter, we conclude the thesis by highlighting the current limitations of the various contributions, and proposing concrete work items that may be considered in order to overcome those limitations as part of future work. We further give an outlook on object class recognition and knowledge transfer from a wider angle, and aim to anticipate future directions on a larger scale.

[2] http://www.mis.tu-darmstadt.de/nlp4vision

RELATED WORK 2

Contents

In this chapter, we give an overview of prior work related to the topic of this thesis, maintaining the distinction between general object class recognition and the more specific field of knowledge transfer, as initiated by Chapter 1. Since the thesis contributes to both fields (Section 1.3), this chapter spans a relatively broad range of prior work. It does by no means provide an exhaustive treatment of all important contributions in either of the fields, but focuses on seminal work and work directly relevant to the constituent chapters of this thesis. Relevance is assumed either as a consequence of inspiration (work that has inspired parts of this thesis) or competition (work that aims at solving similar tasks by different means). Each of the two sections concerned with either of the two fields is subdivided into separate sections for different categories of related prior work, and respective sections that explicitly state the relations between this thesis and prior work, both with respect to commonalities and differences.

2.1 GENERAL OBJECT CLASS RECOGNITION

This section gives an overview of related prior work in the field of general object class recognition. It is divided into separate sections dealing with local image features, shape and perceptual organization, part-based object class models, 3D object class recognition, and probabilistic Markov Chain Monte Carlo inference.

2.1.1 Local features

In this section, we give a brief review of literature related to local image features. The focus lies on comparative studies and evaluations rather than on work actually introducing novel feature types, since this thesis is less concerned with low-level image processing, but, on the other hand, considers higher-level tasks related to general recognition and knowledge transfer. Naturally, the considered high-level tasks depend on lower-level image features, which is why a thorough evaluation of local image features constitutes the starting-point of this thesis (Chapter 3).

Feature evaluations. Mikolajczyk *et al.* (2005b) focus on evaluating affine covariant local region detectors, not taking into account subsequent processing of the obtained regions in terms of feature description. Evaluated detectors comprise Harris points (Harris and Stephens, 1988; Mikolajczyk and Schmid, 2002; Schaffalitzky and Zisserman, 2003), Hessian points (Mikolajczyk and Schmid, 2002), maximally extremal stable regions (Matas *et al.*, 2002), edge-based regions (Tuytelaars and Gool, 1999), intensity extrema (Tuytelaars and Gool, 2000), and salient regions (Kadir *et al.*, 2004). The different detectors are compared in a matching task, using both a measure of repeatability and an actual matching with SIFT (Lowe, 2004) descriptors. The matching is performed on a set of image pairs, where one image of each pair is the result of applying various transformations to the first image, such as viewpoint change, scale change, blur, JPEG compression, and varying illumination. Mikolajczyk *et al.* (2005b) conclude that different detectors are complementary, and thus should be combined for best performance.

Mikolajczyk and Schmid (2005) extends Mikolajczyk *et al.* (2005b) by adding different local feature descriptors to the picture, comprising shape context (Belongie *et al.*, 2000), steerable filters (Freeman and Adelson, 1991), PCA-SIFT (Ke and Sukthankar, 2004), differential invariants (Koenderink and van Doom, 1987), spin images (Lazebnik *et al.*, 2005), SIFT (Lowe, 2004), complex filters (Schaffalitzky and Zisserman, 2002), moment invariants (Gool *et al.*, 1996), cross-correlation, and the newly proposed gradient location orientation histogram (GLOH) descriptor. The evaluation is again performed in the context of matching, but additionally encompasses the recognition of object instances observed under different viewing conditions. Performance is measured in terms of recall and precision. In comparison to Mikolajczyk *et al.* (2005b), rotation is added to the set of examined image transformations. As a result of the evaluation, GLOH and SIFT show the best matching performance, followed by shape context. Hessian point detectors typically perform better than Harris point detectors.

Mikolajczyk *et al.* (2005a) compare the performance of local region detectors and descriptors in the context of object class recognition, using a subset of the detectors and descriptors evaluated by Mikolajczyk *et al.* (2005b) and Mikolajczyk and Schmid (2005), respectively. The evaluation is performed on clusterings of sets of descriptors, constituting a generic, intermediate level of representation that is frequently used at the core of current object class recognition approaches, and which is diagnostic

for inter-class generalization abilities of a given detector-descriptor combination. Generalization ability is measured by cluster precision, and accompanied by an entropy-based measure of feature localization accuracy. Cluster precision experiments are performed on a 20 class subset of Caltech-101. An additional experiment uses the various detectors and descriptors in the pedestrian detection framework of Leibe *et al.* (2005), which is an instance of the implicit shape model (ISM) (Leibe *et al.*, 2006a). Mikolajczyk *et al.* (2005a) conclude that the GLOH descriptor (Mikolajczyk and Schmid, 2005) in combination with Hessian-Laplace points (Mikolajczyk and Schmid, 2002) systematically outperforms all other combinations, followed by salient regions (Kadir *et al.*, 2004). In general, the relative orderings of detector and descriptor performance differ from matching-based evaluations (Mikolajczyk *et al.*, 2005b; Mikolajczyk and Schmid, 2005). The cluster precision results are found to be transferable to object class (pedestrian) detection.

Seemann *et al.* (2005) conduct an evaluation of local region detectors and descriptors, specifically tailored towards pedestrian detection. In comparison to Mikolajczyk *et al.* (2005a), it adds local as well as global Chamfer matching (Gavrila, 2000) to the evaluation. The experimental results show superior recognition performance of shape context descriptors (Belongie *et al.*, 2000) in combination with Hessian-Laplace points (Mikolajczyk and Schmid, 2002).

While Mikolajczyk *et al.* (2005a) focus on planar scenes, Moreels and Perona (2005) compare the performance of local region detector-descriptor combinations on non-flat 3D objects depicted from a variety of different viewpoints. Performance is again measured in the context of image matching. Moreels and Perona (2005) confirm the superior performance of Hessian points (Mikolajczyk and Schmid, 2002) and SIFT (Lowe, 2004) descriptors reported by Mikolajczyk *et al.* (2005a). In general, viewpoint changes of more than 30 degrees tend to deteriorate performance for all detectors and descriptors.

2.1.2 Shape and perceptual organization

In this section, we describe related prior work in the fields of local shape features and perceptual organization. While the first field is typically concerned with object class recognition, the second field comprises work in statistical shape analysis and perceptual grouping, independent of recognition. Both fields are related to this thesis, since most considered object class models are inherently shape-based (chapters 4, 5, 6, 7).

Local shape features. While remarkable recognition performance has lately been reported using local appearance features (Csurka *et al.*, 2004; Fergus *et al.*, 2003; Mikolajczyk *et al.*, 2006), shape-based object class representations are coming into focus again. This recollection of ideas from the earlier days of computer vision is motivated by intermediate advancements in the general design of local features and corresponding progress in statistical machine learning algorithms. The combination of both promises success where the old models often failed, namely, in the robust

matching between model shape and real world imagery.

Ferrari *et al.* (2006b) phrase object class recognition as a graph matching problem. In particular, both a test image and a prototypical shape model of the object class of interest are represented as graphs, and matched by identifying corresponding sub-graphs. The test image graph is termed a Contour Segment Network (CSN), and constitutes a hierarchical representation of a discrete set of contour segments, obtained by application of the Berkeley edge detector (Martin *et al.*, 2004) and successively fitting line segments to image edges. In particular, the CSN connects contour segments according to continuity and proximity conditions, implemented by a set of perceptual grouping rules resembling Gestalt principles. While individual contour segments typically over-segment object boundaries, paths in the contour segment network are likely to correspond to larger portions of or even entire object boundaries. Forming network connections based on the local application of perceptual grouping rules ensures sparse connectivity. Ferrari *et al.* (2006b) introduces a novel data set as a benchmark for shape-based object class recognition, the ETHZ Shape Classes data set. It consists of five object classes characterized by shape rather than appearance, namely, apple logo, bottle, mug, giraffe, and swan. Ferrari *et al.* (2006b) further report promising recognition results on this data set using manually designed prototypical shapes and a greedy graph matching algorithm based on the CSN representation.

Ferrari *et al.* (2008) builds upon the CSN representation of Ferrari *et al.* (2006b), introducing a novel local shape feature based on groups of adjacent contour segments, termed k-AS, where k denotes the size of the considered neighborhood. The corresponding feature descriptor describes the relative spatial layout of its constituents contour segments (relative center distances, relative scales, and rotation angles), and is invariant to translation, scale, and, optionally, rotation. Ferrari *et al.* (2008) gives an extensive evaluation of the k-AS features in object class recognition tasks on a variety of benchmark data sets, in connection with a localized bag-of-features object class detector. In particular, the performance is compared to appearance-based interest point detectors and local feature descriptors, and combinations of different features are explored. k-AS are demonstrated to perform favorably on shape-based object classes compared to appearance-based methods. In particular, 2-AS are shown to offer a good compromise between repeatability and discrimination.

Ferrari *et al.* (2007) combines an object class detector based on k-AS with deformable shape models in the spirit of active shape models (Cootes, 2000). In contrast to (Ferrari *et al.*, 2008), an implicit shape model (Leibe *et al.*, 2006a) is adopted to deliver candidate object detections, which are then verified using the deformable shape model, using thin plate spline robust point matching. An important contribution lies in an iterative shape refinement procedure, which allows learning shape models from real world training images with bounding box annotations. It is capable of separating discriminative object outlines from background clutter, using feature statistics over the training images. The resulting combined object class detector is shown to level the performance of the graph matching-based Contour Segment Network method (Ferrari *et al.*, 2006b) on the ETHZ Shape Classes and Weizmann horses data sets.

Zhu *et al.* (2008) propose contour context selection for recognizing object classes from a single exemplar shape. Recognition is posed as a set-to-set matching problem, and solved using a linear programming relaxation. It comprises a figure-ground labeling of image edges as well as finding set-to-set correspondences between image and model edges. Similarity between shape fragments is measured using shape context descriptors (Belongie *et al.*, 2001). The proposed approach outperforms Ferrari *et al.* (2007) on four of the five classes of the ETHZ Shape Classes data set.

Similar to Zhu *et al.* (2008), Ravishankar *et al.* (2008) propose a shape-based object class recognition system that matches single shape exemplars to real world images. Recognition proceeds in multiple stages, comprising a coarse search for candidate edge fragments, grouping candidate edges, having groups vote for object centroids, identifying maxima in the voting space, and fine-matching hypothesized object contours to image edges using dynamic programming. Experiments are conducted on ETHZ Shape Classes, and demonstrate superior performance compared to Ferrari *et al.* (2007).

The recent work of Srinivasan *et al.* (2010) builds upon the many-to-one contour matching approach proposed by Zhu *et al.* (2008), but replaces exemplar shape matching by discriminative learning from bounding box-annotated training images. The learning problem is phrased as a two stage procedure, in which a representative object class model is first learned from positive training examples, and then further specialized by taking additional negative examples into account. The representative model consists of a set of spatially constrained prototypical contours, formed from image contours by many-to-one matching. Candidate image contours are obtained via a bottom-up grouping process (Zhu *et al.*, 2007). The discriminative specialization is formalized as a latent SVM in the spirit of Felzenszwalb *et al.* (2009), and alternates between hypothesizing contour assignments and updating model parameters accordingly. Detection proceeds in a greedy fashion, by performing local contour searches relative to a dedicated root contour (Felzenszwalb *et al.*, 2009), and uses a linear programming relaxation of many-to-one contour matching. Srinivasan *et al.* (2010) report excellent recognition performance on the ETHZ Shape Classes data set, outperforming most prior work Ferrari *et al.* (2007); Fritz and Schiele (2008); Maji and Malik (2009), and also comparing favorably to the shape-based object class model presented in Chapter 5.

Perceptual organization. While object class recognition is often phrased as a top-down process, in which a prototypical object class template is sought in an image, work in perceptual organization typically aims at explaining image content in a bottom-up fashion. This amounts to identifying potential groupings of local image structures that are non-incidental, e.g., by applying Gestalt principles.

According to Pentland (1986), perceptual organization serves the purpose of recovering structural regularities in visual input data, and making these regularities available as building blocks according to which visual scenes can be (de-)composed. Pentland (1986) develops a computational theory of perceptual organization, in which the formation of visual input data is explained as a succession of generative

and transformative steps, as if "constructing an object from lumps of clay". The corresponding formal description is based on superquadric primitives, boolean combination, and recursive fractal construction.

Brady and Asada (1984) explore a representation of planar shape based on the notion of Smoothed Local Symmetries (SLS). The suggested representation encompasses both the characterization of the boundary of the shape and the region that it occupies. Smoothed local symmetries are defined as maximally smooth assemblies (curves) of local symmetries, where a local symmetry is defined by a specific angular relation between two points on opposing fragments of a given planar shape. The maximally smooth curve is termed the symmetry axis, and often resembles the intuitive notion of a symmetry axis. Brady and Asada (1984) suggest two implementations of SLS. The first one is based on representing planar shapes as a discrete set of points, and exhaustively testing all pairs of points for local symmetry. The second one approximates shapes by primitive line and circle fragments, and by analytically solving for SLS for each pair of primitives. Furthermore, a skeletal decomposition of planar shapes into sub-shapes is formalized by means of SLS, as a replacement for generalized cones as proposed in connection with the ACRONYM system (Brooks *et al.*, 1979).

Saint-Marc *et al.* (1993) describe implementations of skew symmetries, parallel symmetries, and Smoothed Local Symmetries (Brady and Asada, 1984) for the specific case of planar shapes represented as quadratic B-spline curves. For SLS, the complexity of the suggested implementation is quadratic in the number of B-spline curve segments, and based on numerically solving for zero-crossings of a non-linear function.

Zhu (1999) studies a statistical theory of planar shape, focussing on Gestalt laws, such as collinearity, cocircularity, proximity, parallelism, and symmetry. The theory is developed around a Markov Random Field (MRF) model, where shape priors are learned as Gibbs distributions, and neighborhood structures correspond to Gestalt laws. The learned shape models are verified by synthesizing shapes by Markov Chain Monte Carlo (MCMC) sampling, using the Metropolis-Hastings algorithm.

Park *et al.* (2008) give a survey and comparison of various state-of-the-art symmetry detection algorithms, suitable for being applied to real world images. They consider three different symmetry detectors, namely, 1) digital paper cutting (Liu *et al.*, 2005), 2) detecting symmetry and symmetric constellation features (Loy and olof Eklundh, 2006), and 3) detecting rotational symmetries (Prasad and Davis, 2005). 1) uses edge-based local features and a generalized Hough voting procedure to generate candidate reflection axes. 2) supports both reflection and rotational symmetry, and is also based on generalized Hough voting. 3) computes gradient vector flow fields, voting for potential centers of rotational symmetries. The performance of the three symmetry detectors is evaluated both in isolation, on a novel benchmark data set for symmetry detection, and with respect to their potential usefulness for object class recognition. The benchmark data set consists of a collection of synthetic and real world images, both complemented by ground truth symmetry annotations. Performance is measured by sensitivity and false positive rates. For assessing the

potential usefulness of symmetry detection for object class recognition, the benchmark is repeated with images of standard recognition benchmark data sets (PASCAL VOC 2007, Caltech-256, and MSRC), restricted to the object classes of interest. The detected symmetries are not integrated into any object class recognition system. From the experiments, Park *et al.* (2008) conclude that the quality of symmetry detection algorithms is still unsatisfactory, and far from more established feature detectors, such as edge detectors. Nevertheless, they see large potential for symmetry detection to aid object class recognition, motivated by human ability to effectively use symmetry cues, and high success rates of the detectors on individual object classes.

2.1.3 Part-based object class representations

In this section, we give an overview of the most relevant prior work in part-based object class modeling. Part-based approaches are typically favored due to their inherent robustness to partial occlusion, and the often reduced amount of required training data compared to whole-object approaches. The most prominent part-based object class representations comprise the implicit shape model, pictorial structures, and the constellation model. Both the implicit shape model and the constellation model have inspired object class models considered in this thesis (chapters 4 and 5, respectively).

Implicit shape model. Possibly due to its conceptual simplicity yet probabilistic interpretation, the implicit shape model (ISM) has gained remarkable attention in the computer vision literature. Introduced by Leibe *et al.* (2006a), it models objects as flexible arrangements of local features. Local features are represented relative to a pre-trained codebook of visual words, resulting in a vector quantization of the feature space. During training of an object class model, local features collected from training images are matched against the pre-trained codebook, and their location and scale relative to the center of the object of interest is memorized. For recognition, local features are again matched to the codebook, recalling the stored object center locations, and projecting them as object hypotheses to the test image plane, which resembles a generalized Hough voting procedure. The voting space is typically three-dimensional, consisting of object location and scale. Applying kernel-based density estimation techniques, the modes of the resulting distribution of votes can be output as the final detection hypotheses. Leibe *et al.* (2006a) propose an additional verification step based on segmentation mask fragments stored along with each local feature and the principle of minimum description length (MDL), which improves recognition performance in particular for multiple object class instances in cluttered scenes. The implicit shape model has since been used in a variety of contexts, and shown to deliver excellent recognition performance for general object class recognition (Leibe *et al.*, 2006a), pedestrian detection (Seemann *et al.*, 2007), and part-based people tracking-by-detection (Andriluka *et al.*, 2008).

Maji and Malik (2009) augment the implicit shape model by an additional discriminative training step, which assigns high weights to features that provide reliable separation between positive and negative training examples. The weights depend on both the appearance of individual features and their combined spatial distribution, and can be optimized using standard convex optimization techniques in a maximum margin framework. The resulting discriminatively trained ISM is further combined with an additional discriminative verification stage (Lazebnik *et al.*, 2006), and shown to outperform the purely representative ISM variant on various data sets (ETHZ Shape Classes, UIUC Cars, INRIA Horses).

Ommer and Malik (2009) propose a refinement to the original ISM formulation that circumvents relying on often unreliable scale estimates provided by local feature detectors. Their formulation is based on the observation that the scale of a local feature is often not uniquely determined from its descriptor due to an inherent scale-location ambiguity, and consequently lets features vote for entire ranges (lines) in scale space rather than for individual points. As a result, voting is implemented by agglomerative clustering of lines in scale space rather than by accumulation of point votes and subsequent density estimation. Experimental results are reported on the ETHZ Shape Classes data set, and demonstrate the superiority of the proposed over the standard ISM formulation. Similar to Maji and Malik (2009), an additional discriminative verification stage (Lazebnik *et al.*, 2006) using bootstrapping is shown to further improve performance.

Pictorial structures. The pictorial structures model adds representational power to the implicit shape model, while still maintaining computational tractability during recognition. While the probabilistic interpretation of the spatial layout component of the ISM corresponds to a star-shaped topology, where each node depends on a unique reference node, pictorial structures models correspond to arbitrary tree structures (i.e., cycle-free graphs). This allows the formulation of tighter spatial constraints between parts than with the ISM, resulting in potentially more discriminative object class models. Following the initial idea of spring-like connections between object parts presented by Fischler and Elschlager (1973), Felzenszwalb and Huttenlocher (2000) give a probabilistic formulation of pictorial structures and an efficient maximum a posteriori (MAP) algorithm for matching pictorial structures models to test images for recognition. The efficient MAP procedure is based on dynamic programming and generalized distance transforms.

Because of its tree structure, the pictorial structures model lends itself to representing articulate object classes, such as humans. Human body parts, such as the torso, arms, and legs, can be mapped one-to-one to model parts, and characterized by permissible angular constraints at the joints between adjacent parts. Andriluka *et al.* (2009) demonstrates state-of-the-art performance for people detection and tracking, using discriminatively trained part detectors (dense shape context (Belongie *et al.*, 2001) in connection with AdaBoost (Freund and Schapire, 1997)) and a GPLVM prior of the human walking cycle.

Recently, the pictorial structures model has also been revisited for general object

class recognition. Felzenszwalb *et al.* (2009) propose to train a mixture of multiple pictorial structures models in a discriminative fashion, using the formalism of latent support vector machines (SVMs). The mixture representation allows to capture multiple aspects of an object class, such as different viewpoints, in one coherent model, while the latent SVM formulation allows for weakly supervised training by approximate bounding boxes. An individual aspect-level pictorial structures model is defined relative to a root filter, which amounts to a part detector trained on entire objects. The underlying visual feature representation borrows from histograms of oriented gradients (HOG) proposed by Dalal and Triggs (2005). The given experimental evaluation suggests excellent recognition performance on a variety of object classes of the PASCAL VOC challenges.

Constellation model. While pictorial structures models sacrifice representational power in favor of efficient inferencing, the class of constellation models does not pose any restrictions on the structural dependencies between parts of a given object class model, allowing fully connected dependency graphs. As a consequence, exact inference involves enumerating all possible assignments between model parts and image features, limiting the number of features that can be considered during recognition. The constellation model evolved over a series of works by Michael C. Burl and colleagues, until it was popularized by Fergus *et al.* (2003). Departing from an initial variant supporting single-scale recognition of parts of fixed appearance (Burl and Perona, 1996), Burl *et al.* (1998) added soft-assigned part detections and scale invariance to the model. Weber *et al.* (2000) replaced manually provided part annotations in training images by unsupervised learning using expectation maximization (Dempster *et al.*, 1977).

The main contribution of Fergus *et al.* (2003) consists in the explicit integration of part appearance variability into the model, and extending the learning procedure accordingly, again using expectation maximization. Fergus *et al.* (2003) model part appearance as Gaussian densities over PCA-projected image patches, extracted from salient regions (Kadir *et al.*, 2004) at multiple scales. Common to all variants of the constellation model is the representation of the goodness of fit between model and image as a likelihood ratio between foreground and background hypotheses, respectively. This discriminative formulation is particularly important for two reasons. First, it allows to separate relevant object class features from randomly co-occurring background clutter for unsupervised learning of parts. Second, it allows to model missing parts by assuming they have been generated by a background density. Recognition is phrased as image-level classification, integrating over all possible hypotheses (assignments between image and model features). The exhaustive search for hypotheses is sped up by either organizing the search process (A-star search) such that previous computations can be reused, or hypotheses are rejected early due to bound computations and thresholding (Fergus *et al.*, 2001). Since the exhaustive search is also part of the unsupervised learning procedure, Helmer and Lowe (2004) suggest to learn constellation models incrementally, by successively adding parts to yield increasingly complex models. Due to its powerful statistical model, the

constellation model incarnation of Fergus *et al.* (2003) exhibits superior recognition performance compared to prior work on the Caltech-4 data set. Fergus *et al.* (2004) extend the constellation model to support heterogeneous parts, by combining image patches and curve segments. The model is shown to benefit from the complementary nature of the parts in an image search re-ranking task. Extensions to the constellation model proposed in the context of transfer learning (Fei-Fei *et al.*, 2006) and multi-view recognition (Savarese and Fei-Fei, 2007, 2008; Sun *et al.*, 2009; Su *et al.*, 2009) are discussed in the respective specialized subsections of this chapter.

2.1.4 3D Object class recognition

This section is concerned with three-dimensional object class representations in the broadest sense. While, in the early days of computer vision, these representations were predominant, more recent approaches use 3D object class models mostly in connection with multi-view recognition, due to the inherent three-dimensional nature of the problem. 3D object class recognition is related to this thesis, since it proposes an object class model that can be learned from 3D CAD models, and used to detect objects from multiple viewpoints (Chapter 7). In the following, we distinguish among early approaches, implicit shape model-based approaches, probabilistic generative models, partial surface models, models based on 3D surface meshes, and a specific class of image features related to shading artifacts.

Early approaches. Nevatia and Binford (1977) propose a method to recognize curved 3D objects in laser range data. It is phrased as a three step procedure, consisting of segmenting the range image into a collection of generalized cone primitives (GC), describing the topology of the obtained segmentation by symbols, and matching the symbolic description to a database of stored exemplar descriptions. Symbolic descriptions are graph-structured, connecting individual segments by joints, and contain quantitative information on geometric properties of both segments and joints. Recognition then amounts to graph matching, implemented as a greedy procedure that matches increasingly complex sub-graphs, starting from a set of seed matches (which can be efficiently obtained through indexing). The method is limited to instance recognition of simplistic objects (toy horse, hammer) and scenes.

Marr and Nishihara (1978) give a more general and theoretical treatment of desirable characteristics of shape representations suitable for recognition, but arrive at a representation similar to Nevatia and Binford (1977). Following the three criteria accessibility, scope and uniqueness, and stability and sensitivity, the proposed representation is object-centric, modular, and based on volumetric primitives. In particular, hierarchical compositions of cylindrical shapes are considered and described by the characteristic geometric relations between their respective canonical axes. Recognition proceeds by first obtaining a canonical axis description from two-dimensional image data, and consequently matching this description to a database of stored exemplar descriptions, using various indexes. The mapping between recognized 2D and stored 3D descriptions is achieved by focusing on topological aspects of shape

that are invariant to projection.

In contrast to this, Brooks *et al.* (1979) explicitly incorporate projective constraints into the ACRONYM system, in order to match 3D generalized cylinder descriptions to 2D images. The system is based on a geometric reasoning engine, and recognizes objects by a succession of prediction, description, and interpretation steps. All three steps are represented as graphs, generating an interpretation graph from matching the graphs corresponding to prediction and description, respectively. Connected components of the interpretation graph correspond to detection hypotheses. The system has been applied to airplane detection in aerial images.

Similar in spirit to (Marr and Nishihara, 1978), Lowe (1987) chooses projection invariant structures as the starting point for model-based recognition, but also employs a viewpoint consistency constraint in order to simultaneously solve for viewpoint and model parameters. The model is represented as a set of pairs of 3D straight line segments and visibility flags for projection. Projection invariant matching candidates are obtained by grouping image line segments according to perceptual organization principles (proximity, parallelism, and collinearity). Recognition starts from an initial set of matches between image and model line segments, and tentatively refines projection as well as model parameters for each match using Newton's method. In principle, the proposed approach is capable of handling cluttered backgrounds and occlusion of image line segments. However, it is limited to instance recognition, and relies on 3D model line segments giving rise to corresponding line segments in 2D images.

Implicit shape models. A whole line of research has been devoted to extending the implicit shape model (ISM) (Leibe *et al.*, 2006a) to multi-view recognition. Thomas *et al.* (2006) propose to share features among multiple ISMs (one per viewpoint) through activation links, connecting codebook entries of neighboring viewpoints. Activation links are obtained by tracking features across training images of multiple viewpoints (Ferrari *et al.*, 2004), and traversed during recognition in order to benefit from the accumulated evidence of multiple training viewpoints. Performance is evaluated on the PASCAL VOC 2005 data set (object classes motorbike and sports shoe), and shown to be superior to a bank of detectors model not using activation links. Similar in spirit, but different in implementation, Arie-Nachimson and Basri (2009) establish correspondences between features of different viewpoints by means of similarity transforms. Furthermore, their work subsumes all training image features in a single, three-dimensional ISM, which is built in a two stage procedure. First, an initial approximate 3D model is constructed from a dense collection of uncalibrated camera images of a single object class instance. This instance specific model is then enhanced with training images of additional object class instances, by matching features in the vein of 2D ISMs. Recognition is performed in a RANSAC (Fischler and Bolles, 1981) loop, interleaving the estimation of model likelihood and image projection. The paper reports moderate recognition and viewpoint classification performance for the PASCAL VOC 2007 and 3D object classes (Savarese and Fei-Fei, 2007) car data sets. Yan *et al.* (2007) also construct a 3D ISM by reconstructing a

model from training images, but using a homography framework (Khan *et al.*, 2007). Additional training image features are attached based on appearance similarity, as measured between their SIFT (Lowe, 2004) descriptors. The model is shown to perform moderately on PASCAL VOC 2006 motorbikes, and comparable to related work on horses.

Generative models. Another line of research uses probabilistic generative models to capture object class appearance and geometry under varying viewpoints. Although not fully generative in nature, Savarese and Fei-Fei (2007) marks the beginning of a series of papers following this principle direction, and introduces a new benchmark data set for multi-view recognition (the 3D object classes data set). The object class model of Savarese and Fei-Fei (2007) can be interpreted as an extension of Weber *et al.* (2000) and Fergus *et al.* (2003) to the multi-view case, where object parts are not only subject to probabilistic spatial constraints, but additionally linked between viewpoints by homographies. The model is evaluated on the newly proposed multi-view data set with respect to object class recognition and viewpoint classification, and demonstrated to outperform Thomas *et al.* (2006) on PASCAL VOC 2005 motorbikes. Savarese and Fei-Fei (2008) extends that model by synthesizing previously unseen views of object classes at recognition time. View synthesis is achieved by interpolating each of the appearance of individual parts, their spatial layout, and their homographic linkage structure, between up to three views available during training. As a result, Savarese and Fei-Fei (2008) consistently outperforms Savarese and Fei-Fei (2007) on the 3D object classes data set. Sun *et al.* (2009) gives a rigorous generative formulation of the former model, phrased as a Dirichlet process Gaussian mixture. The model is learned using expectation maximization (Dempster *et al.*, 1977). Interestingly, the learned generative model is not employed as is for recognition. Instead, the learned part appearance distributions are replaced by sliding window part detectors, and the learned spatial part layout constraints are turned into a generalized Hough voting procedure in the spirit of the ISM (Leibe *et al.*, 2006a). The generative model outperforms Savarese and Fei-Fei (2008) on 3D object classes cars, and achieves modest recognition performance on PASCAL VOC 2006 cars. By incorporating view synthesis (Savarese and Fei-Fei, 2008) into the generative formulation (Sun *et al.*, 2009), Su *et al.* (2009) further improve the performance on 3D object classes cars and bicycles. In particular, they propose a triangularization of the viewing sphere, allowing to synthesize viewpoints not part of the triangularization. As a result, the learning of multi-view object class models can be bootstrapped from a video sequence obtained by walking around an instance of the object class.

Partial surface models. Kushal *et al.* (2007) propose collections of partial surface models (PSMs) for part-based recognition of 3D objects. PSMs constitute dense, locally rigid assemblies of texture patches, that can be robustly matched between training images of an object class, using affine transformations. A distinct PSM is allocated to each distinct view of an object part, and the individual PSMs are

linked in a graph describing their relative spatial layout. Individual PSMs are based on logistic regression classifiers. The graph structure is learned using an iterative procedure and loopy belief propagation (Murphy *et al.*, 1999). Object recognition approximates a MAP solution by a greedy local search algorithm. Experiments are conducted on PASCAL VOC 2005 cars, and performance is demonstrated to exceed that of related work.

The more recent approach of Gill and Levine (2009) is similar in spirit to PSMs, in that it models object parts as collections of dense SIFT (Lowe, 2004) descriptors, and imposes constraints on their relative layout. However, the appearance representation is based on local linear embedding (LLE), and spatial constraints are enforced at recognition time using a greedy search strategy. Gill and Levine (2009) achieve excellent recognition performance on the 3D object classes data set, outperforming Su *et al.* (2009).

3D Surface mesh-based models. The limited accuracy with which 3D geometry can be estimated from scarce real world training images, in particular on the level of object classes, naturally leads to approaches that employ existing high-quality 3D meshes as a source of information. Liebelt *et al.* (2008) propose to learn 3D ISM (Leibe *et al.*, 2006a) models from collections of fully textured 3D meshes, which are rendered photorealistically from a variety of viewpoints and in front of varying backgrounds, in order to resemble real world image statistics as closely as possible. Standard local feature detectors and descriptors (Bay *et al.*, 2008) are then applied to the rendered images, and used to populate an appearance codebook. As demonstrated in their experiments, performance depends crucially on a discriminative filtering step, which consists of two components. First, codebook features are weighted according to their repeatability against varying backgrounds. Second, an SVM classifier is trained to distinguish rendered from real world image features, and subsequently used to prune a large fraction of image features prior to object pose voting. While the 3D ISM provides a rough estimate of object pose, that estimate is refined using RANSAC (Dempster *et al.*, 1977) and a perspective three-point method (Haralick *et al.*, 1994). Performance is evaluated on PASCAL VOC 2006 cars and motorbikes, and shown to be competitive with state-of-the-art 2D object class detectors. In particular, for motorbikes, precision drops much more slowly for increased recall in comparison to other methods, underlining the discriminative power of the model.

Departing from the exclusive use of 3D models in Liebelt *et al.* (2008), the more recent work of Liebelt and Schmid (2010) treats appearance and geometry as two completely separate learning tasks. While geometry is again learned using 3D meshes, appearance is learned from real world instead of rendered images. While the authors expect increased robustness of the resulting appearance representations, separating the two tasks during learning asks for an explicit mapping between them for recognition. The mapping is achieved by rendering 3D models from the same viewpoints found in the training images, and overlaying both with the same regular grid, establishing correspondences between respective grid positions. Consequently, 2D image positions are associated 3D mesh positions, and vice versa. The chosen

appearance representation is a combination of dense image features (Tola *et al.*, 2010) and spatial pyramids (Lazebnik *et al.*, 2006), and used on both the level of whole objects and object parts, as defined by cells of the above described regular grids. The geometry representation models the occupancy of corresponding grid volumes as 3D Gaussian mixtures. Recognition is a three phase procedure, in which regions of interest are predetected, and base viewpoints are predicted from part detector responses, using a voting scheme. Then, 3D object pose and camera parameters are simultaneously estimated using expectation maximization (Dempster *et al.*, 1977). The proposed method is shown to outperform related work (Su *et al.*, 2009; Gill and Levine, 2009) on 3D object classes cars, and to perform comparable to related work on bicycles.

Shading cues. Brightness variations on illuminated surfaces constitute powerful cues for the human perception of three dimensional shape (Kleffner and Ramachandran, 1992; Koenderink *et al.*, 1996). Unfortunately, the general shape-from-shading problem, i.e., estimating the geometry of three-dimensional surfaces from apparent shading artifacts in real world images, is considered difficult mostly due to its inherent ambiguities (Horn and Brooks, 1989), leading to heavily underconstraint problem formulations. While this can in principle be accounted for by regularization, the price is reduced closeness to the originating image data. As a consequence, off-the-shelf shape-from-shading techniques are rarely used in object class recognition. Nevertheless, the potential to directly hint on three-dimensional object shape have lead to numerous attempts to exploit shading artifacts for recognition.

Following psychological evidence that humans can perceive relative depth more accurately than absolute depth from shading, Weinshall (1992) propose an approximate local shape-from-shading technique. While this technique does not provide an exact global estimate of surface normals, it allows to infer relative depth values in local regions near global shading maxima. In particular, it allows to classify these regions into parabolic, elliptic, and hyperbolic shapes.

Haddon and Forsyth (1998) concentrates on the characterization and recognition of shading primitives, visual artifacts in real world images which are strongly coupled to surface shape. The work focuses on cylindrical surfaces as one possible instance of shading primitives, motivated by their ubiquitous appearance, such as in human or animal limbs. The approach is based on training a support vector machine (SVM) classifier from synthetic data, generated from a geometric model of roughly cylindrical shapes, and a shading model. The shading model assumes surface brightness being composed of two components, distant radiation, modeling the effect of diffuse interreflections between distant surfaces, and a single point light source at infinity. Positive training data for limbs is obtained by PCA projections of rendered cylinder profile images. Negative data is taken from randomly sampled lines in background images. The learned limbness classifier is used to verify limb hypotheses generated by edge detection and subsequent grouping of edges according to symmetry. In a second experiment, the limb classifier is extended to detect human backs, which involves identification of its center groove, artificially filling in the

groove, and applying the original limb classifier to the filled-in candidate region.

Worthington and Hancock (2001) explore the usefulness of three-dimensional surface topography information acquired from shape-from-shading for object (instance) recognition from two-dimensional images. For this purpose, two directions are pursued. The first uses a histogram representation from shape-from-shading needle maps in connection with a nearest neighbor classification scheme. The second uses a mid-level abstraction of needle maps (constant shape-index maximal patches), subsumed in a graph of adjacent regions, used in a matching framework. Experiments are conducted on the COIL data set (Nene *et al.*, 1996) and reveal promising recognition performance.

The goal of Mori *et al.* (2004) is to estimate human body pose from single real world images, using a probabilistic model combining various visual features and a prior distribution on plausible body poses. The underlying image representation is based on an image over-segmentation by superpixels. Besides the shape of adjacent superpixel configurations, shading artifacts are considered as a "limbness" measure (Haddon and Forsyth, 1998). Instead of explicitly applying a shape-from-shading model, Mori *et al.* (2004) resort to a learning-based approach, which represents limb shading by a prototypical intensity template. The cross-correlation between template and image content is used as a limbness measure.

Lichtenauer *et al.* (2005) compute isophotes (lines of equal brightness) in real images, and use their direction and curvature as features for face recognition. Isophote computation is performed using an orientation tensor representation. Isophote orientation is split into magnitude and sign components, and used in connection with various classifiers for recognition. Experiments on a face data set confirm the superiority of the proposed isophote features compared to intensity-, gradient-based, and Haar-like features.

Wu *et al.* (2007) propose to base facial gender classification on a 2.5 dimensional shape representation (needle maps) obtained via shape-from-shading. The proposed method interleaves the construction of needle maps with the estimation of a statistical shape model, based on principal geodesic analysis, which balances image evidence and model regularization. Linear discriminant analysis (LDA) is then applied to learn discriminative models for either gender. Experiments are conducted on a database of manually registered face images, and the proposed method is shown to yield gender classification performance on par with human performance.

Nillius *et al.* (2008) introduce a set of generic detectors for three dimensional shape primitives, based on shading artifacts in real world images, similar in spirit to Haddon and Forsyth (1998). The approach combines a physically plausible model with learning, by model-based principle component analysis (PCA). In particular, principle components are determined analytically from the physical model, and not estimated from data. The physical model assumes a single distant light source, illuminating a scene observed by orthographic projection. The bidirectional reflectance distribution function (BRDF) is modeled in frequency space using a spherical harmonics decomposition, and not restricted to Lambertian reflectance. The approach is instantiated for the detection of spheres and cylinders, and applied in a sliding

window fashion at multiple scales and rotations in test images. An estimate of the residual error between image evidence and a PCA-based reconstruction yields a measure of detection confidence. Experiments are conducted on both synthetic and real world images, and demonstrate the effectiveness of the proposed shading primitive detectors.

2.1.5 Markov Chain Monte Carlo inference

Approximate probabilistic inferencing schemes based on Markov Chain Monte Carlo methods (Gilks *et al.*, 1996) have been employed in various contexts in computer vision, in case exact inference has proven intractable. Among those methods, data-driven (DDMCMC) techniques are of particular interest, since these allow to guide the sampling process by bottom-up proposals, resulting in a more effective visitation of high density regions of the search space. DDMCMC methods are related to this thesis, since Chapter 5 introduces an efficient MAP approximation for the constellation model based on DDMCMC techniques.

Scene interpretation. Following this argumentation, Zhu *et al.* (2000) propose a principled, generic framework for object recognition, combining bottom-up cues with top-down verification. While their approach is demonstrated mostly on synthetic data, it encompasses the combination of different heterogeneous bottom-up cues, and models both objects and scenes in a coherent, structured representation. In particular, scenes can be composed of a variable number of objects. The corresponding state space of the Markov Chain thus has sub-spaces of different dimensionality, and transitions between these sub-spaces are implemented as reversible, trans-dimensional jump moves (Green, 1995), using the Metropolis-Hastings algorithm (Gilks *et al.*, 1996).

Tu *et al.* (2005) instantiate the framework established by Zhu *et al.* (2000) for parsing images into their constituent sub-structures, such as shading or texture patterns, and object regions. The goal of image parsing is the construction of a parsing graph for a given image, where each nodes explains part of the image content. The parsing graph is hierarchical, and combines a top-down composition of the image with information about the relative geometric layout of image regions. The construction of the graph is phrased as a DDMCMC procedure, in which discriminative bottom-up cues guide a sampling process towards likely states of a generative model linked to the graph. In particular, bottom-up cues comprise region boundary cues, face region cues (obtained from a discriminatively trained detector (Viola and Jones, 2001)), text region cues, and cues characterizing pairwise affinities between image regions. During sampling, the parsing graph is constantly modified by adding new components (birth move), deleting existing components (death move), merging multiple components (merge move) or splitting them (split move). Image parsing is demonstrated on a novel data set of street scenes, comprising face detection, text detection, and image segmentation tasks.

Wojek *et al.* (2010) follow a similar approach, specifically tailored towards visual scene understanding for mobile platforms, integrating specialized detectors for various object classes (pedestrians, vehicles) with scene segmentation, horizon estimation, and various other sensorical measurements. The resulting system achieves state-of-the-art performance on challenging street scene sequences, by additionally enforcing temporal coherence among subsequent frames.

Human body pose estimation. In the context of estimating human body poses from still images, Lee and Cohen (2004) implement the DDMCMC paradigm by using proposal maps to guide three-dimensional pose search. Proposal maps are generated from a variety of cues, comprising face detection, head-shoulder contour matching, elliptical skin blob detection, and detection of ridges. Human body poses are scored by a likelihood function based on a human kinetics model, a human body shape model, a clothing model, and combined with a prior distribution over plausible poses.

2.1.6 Relation to own work

In this section, we highlight the relations between the work presented in this thesis and related work, for each individual chapter, as concerns general object class recognition. The corresponding relations for the more specific case of knowledge transfer are subject of Section 2.2.4.

Chapter 3: Local features for classes of geometric objects. The feature evaluation conducted in Chapter 3 is inspired by Mikolajczyk *et al.* (2005a). It is targeted towards the specific task of object class recognition, and chooses clusterings of local features as the starting-point of the analysis, following the argumentation of Mikolajczyk *et al.* (2005a). In contrast to Mikolajczyk *et al.* (2005a) however, its focus lies on shape, with respect to both the local features and the test object classes under consideration. In particular, it adds the shape-based shape context (Belongie *et al.*, 2000), geometric blur (Berg and Malik, 2001) and k-AS (Ferrari *et al.*, 2008) to the evaluation, and proposes two novel data sets (*Shape* and *Shape2*) of object classes that are characterized by shape rather than appearance. Chapter 3 explores different clustering algorithms, and suggests an improvement over the previously proposed cluster precision measure (Mikolajczyk *et al.*, 2005a). The clustering-level experiments are transfered to a proper recognition setting on both the shape-based data sets and various subsets of Caltech-101, using (localized) bag-of-words representations in connection with a variety of different classifiers, linking to the evaluation of k-AS features provided by Ferrari *et al.* (2008).

Chapter 4: Functional object class detection. The object class model at the core of the approach presented in Chapter 4 is an instance of the implicit shape model (ISM) (Leibe *et al.*, 2006a), in combination with k-AS features (Ferrari *et al.*, 2008), which exhibit excellent performance in the evaluation of local features in Chapter 3.

Relevant visual features are selected from video sequences, which are pre-processed by a foreground-background segmentation technique (background cut (Sun *et al.*, 2006)) based on a conditional random field (CRF) (Lafferty *et al.*, 2001) and pre-trained skin color models (Jones and Rehg, 1999), implemented using graph cuts (Boykov and Kolmogorov, 2004).

Chapter 5: Shape-based object class model for knowledge transfer. Concerning general object recognition, the work presented in Chapter 5 is inspired by the constellation model (Burl *et al.*, 1998; Fergus *et al.*, 2003), in that is part-based, and describes plausible part configurations in a probabilistic fashion. In contrast to prior work (Fergus *et al.*, 2001, 2003; Helmer and Lowe, 2004) based on exhaustive search, we suggest data-driven Markov Chain Monte Carlo (DDMCMC) sampling for efficient approximate MAP inference, allowing to scale the number of features that can be processed per image from dozens (Fergus *et al.*, 2003) to thousands. While DDMCMC methods have been successfully applied to scene interpretation (Tu *et al.*, 2005; Wojek *et al.*, 2010), it has rarely been applied to general object class recognition with part-based models, with the notable exception of human body pose estimation (Lee and Cohen, 2004). Also in contrast to prior work, our approach is entirely shape-based, extending the *k*-AS features of Ferrari *et al.* (2008) by a B-spline curve representation. The descriptive power of pair-wise relations between features has previously been demonstrated by Leordeanu *et al.* (2007), and motivates the introduction of pair-wise symmetry relations between object parts. The concrete flavor of symmetry features is reminiscent of early shape-based approaches (Brady and Asada, 1984), and realized by a B-spline implementation (Saint-Marc *et al.*, 1993). To our knowledge, our work is the first to exploit symmetries in shape-based object class recognition (Park *et al.*, 2008), and demonstrate their usefulness on a standard benchmark data set (ETHZ Shape Classes (Ferrari *et al.*, 2006b)), outperforming two related approaches at the time of publication (Ferrari *et al.*, 2007; Fritz and Schiele, 2008).

Chapter 6: Shading cues for object class detection. Since the work presented in Chapter 6 extends the object class model introduced in Chapter 5, it generally inherits the relations to prior work. In addition to those, it is inspired by the shading primitives of Weinshall (1992), Mori *et al.* (2004), and Haddon and Forsyth (1998), and adheres to a similar hypothesis verification paradigm, in which a 3D shape hypothesis is verified by shading cues. The underlying shading model, assuming Lambertian surfaces, constant ambient lighting and a single distant point light source, is similar to the ones proposed by Weinshall (1992); Haddon and Forsyth (1998). In contrast to prior work, the resulting combined object class model is shown to benefit from shading information on a standard recognition benchmark (Ferrari *et al.*, 2006b), and successfully handles non-Lambertian surfaces by discarding specular highlights as outliers (Fischler and Bolles, 1981). Similar in spirit to Nillius *et al.* (2008), we propose shading cues based on a physical model. The approach of Nillius *et al.* (2008) differs from ours in that it generates bottom-up shading primitive candidates,

requiring a costly sliding-window search over image locations and scales. Ours is inherently top-down, avoiding dense computation of isophotes (Lichtenauer *et al.*, 2005) and surface normals (Wu *et al.*, 2007) altogether. In contrast to Nillius *et al.* (2008), which is capable of handling cylindrical as well as spherical shapes, our current implementation is limited to cylindrical shapes.

Chapter 7: Learning shape models from 3D CAD data. The multi-view recognition approach of Chapter 7 follows the bank-of-detectors paradigm coined by Thomas *et al.* (2006). Each individual detector combines part-detectors based on shape context features (Belongie *et al.*, 2001) that have proven successful in the context of object class recognition (Andriluka *et al.*, 2009) with the powerful spatial model introduced in Chapter 5. This powerful spatial model stands in contrast to most prior work in multi-view recognition, which almost exclusively resorts to a star-shaped model for recognition, using generalized Hough voting (Thomas *et al.*, 2006; Yan *et al.*, 2007; Liebelt *et al.*, 2008; Su *et al.*, 2009; Arie-Nachimson and Basri, 2009).

Chapter 8: Semantic relatedness for knowledge transfer. The work of Chapter 8 builds upon well-established grounds in object class recognition. In particular, it uses various feature channels, such as SIFT (Lowe, 2004), rgSIFT (van de Sande *et al.*, 2010), PHOG (Bosch *et al.*, 2007), SURF (Bay *et al.*, 2008), and local self-similarity histograms (Shechtman and Irani, 2007), in connection with spatially constrained bag-of-words representations (Csurka *et al.*, 2004) and support vector machine (SVM) classifiers. Performance is evaluated on the publicly available Animals-with-Attributes (AwA) data set, using the available pre-computed features (Lampert *et al.*, 2009). The evaluation goes beyond the one given in Lampert *et al.* (2009) by additionally exploring a more realistic zero-shot classification task in which training and test object classes can not be assumed disjoint.

2.2 KNOWLEDGE TRANSFER

This section gives an overview of prior work related to knowledge transfer in object class recognition and, more generally, in computer vision. It distinguishes among knowledge transfer that is driven purely by visual information (Section 2.2.1) from knowledge transfer using additional sources of information (Section 2.2.2), such as language resources. The distinction between the two different flavors of transfer is not always well defined. Assuming and exploiting a given hierarchical structure on the space of object classes, for instance, clearly constitutes knowledge beyond purely visual information. On the other hand, object class labels are naturally given as part of every supervised classification task, and hence not assumed additional sources of information. A third variant of knowledge transfer, although seldom made explicit in the corresponding literature, comprises the generalization of object classes beyond basic-level object categories (Rosch *et al.*, 1976), often in the form of functional or

affordance-based (Gibson, 1977) categorization (Section 2.2.3).

2.2.1 Visual knowledge transfer

Knowledge transfer purely based on visual information can be roughly categorized into the joint learning of multiple object classes, distance-based models, attribute-based models, approaches that exploit object class hierarchy, and Bayesian priors. Distance- and attribute-based models as well as the use of Bayesian priors have inspired the works presented in chapters 8 and 5, respectively.

Joint learning of multiple object classes. Motivated by the human ability to benefit from commonalities of multiple learning tasks posed simultaneously, the notion of multiple task learning (MTL) has been reflected in the machine learning literature. Ben-David and Schuller (2003) are among the first to give a solid theoretical analysis of MTL, based on a data generation model for related tasks. While the proposed model is limited to certain classes of learning tasks, it provides analytical upper bounds for the number of needed training samples. In particular, it gives a formal definition of the relatedness between different learning tasks, based on transformations between the feature spaces associated to each of the individual learning tasks. In this model, multiple task learning is proven to require less training examples than single task learning.

Torralba *et al.* (2004) apply the idea of multiple task learning to multi-class classification, implemented as a variant of the classical boosting algorithm (Freund and Schapire, 1997), termed joint boosting. By sharing features (weak classifiers) across object classes, joint boosting effectively reduces both computational complexity and required training data for recognition. In particular, the number of required features for a given performance level grows only sublinearly in the number of object classes. Feature sharing is demonstrated to outperform classical boosting on a variety of data sets, and shown to be effective in multi-view recognition. Experiments confirm the intuition that neighboring views share visual features.

Amit *et al.* (2007) advocate the joint learning of object classes and a set of shared, latent characteristics. These characteristics are expressed as linear transformations on the input space, equivalent to a latent feature representation. In connection with a linear multi-class classification framework, learning can be phrased as a joint convex optimization problem. This still holds true for kernelized multi-class classification. Experiments on both handwritten character data and the Mammals data set (Fink and Ullman, 2008) show improved recognition performance for classes for which few training examples are available.

Distance-based models. Representing previously unseen object classes by means of their distances to already known classes in some feature space has been adopted by several approaches, often in connection with learning object class models from few training examples. Fink (2004) propose to learn a generic class relevance pseudo metric for 1-shot recognition. Being based on a kernel representation, this pseudo

metric is trained to attain larger values when comparing two instances of different object classes than when comparing instances of the same class. A large margin criterion selects a subset of relevant feature dimensions, which is assumed to be transferable to previously unseen object classes. Single-shot recognition is performed following a 1-nearest neighbor scheme, projecting both training and test data to the selected relevant feature dimensions. The validity of the method is demonstrated on both synthetic and real world character data.

Bart and Ullman (2005b) represent a previously unseen object class as a vector of distances, measured between a prototypical instance of that class and instances of a set of known reference classes. For recognition, the distance vector is computed for the test instance, and classified using a 1-nearest neighbor scheme, termed cross-generalization. While Bart and Ullman (2005b) suggest to compute distances from an image fragment-based visual feature representation, the proposed approach can in principle be applied in connection with any set of classifiers providing confidence values (which can be interpreted as inverse distances). Experimental results are reported on a variant of Caltech-101, and demonstrate the effectiveness of the proposed cross-generalization scheme compared to a standard classifier baseline. Similar to Fink (2004), Bart and Ullman (2005a) propose a feature selection mechanism (again termed cross-generalization) that emphasizes useful features for a known set of classes when recognizing previously unseen classes. In addition, the formerly useful features are adapted to the novel classes. Specifically, each formerly useful feature is replaced by a corresponding feature of a single training example of a novel class, determined by the distance between the two features. Experimental results are reported on the data set introduced by Bart and Ullman (2005b), and confirm the validity of the proposed method.

Thrun (1996) formulates the re-use of once acquired knowledge as a lifelong learning problem, in which the n-th learning task potentially benefits from earlier processed $n-1$ learning tasks. Using the example of memory-based learning, such as nearest neighbor schemes, Thrun (1996) explores two alternative routes. The first one is based on learning a new data representation for task n from the training data of tasks 1 to $n-1$, minimizing a distance criterion in the spirit of Fink (2004). The second one is based on learning a generalized distance function from the training data of tasks 1 to $n-1$, which can determine whether any two instances belong to the same concept. Both routes are implemented as extensions to standard neural network back-propagation, in the form of learning with hints and explanation-based neural network learning, respectively. Experiments on a novel image data base confirm improved generalization abilities of the lifelong learning models, in particular for scarce training data.

Attribute-based models. Characterizing objects according to descriptive attributes has attained increasing attention in recent literature. Attributes offer the benefit of encoding high-level visual properties that can be potentially be shared and reused among several object classes, hence promoting scalability of recognition.

Ferrari and Zisserman (2007) were among the first to pursue this direction, by proposing a generative model for visual attributes. Being built on image segments, their notion of attributes encompasses both unary properties, such as segment color or shape, and binary relations between them, such as relative geometric layout. The model can be learnt in a weakly supervised, discriminative fashion (the presence or absence of an attribute in a set training images is sufficient). An approximate, iterative procedure maximizes the ratio between the likelihoods of image segments being generated by foreground attributes and background, respectively. Learned attribute models can then localize attributes on the level of individual or pairs of segments in unseen test images.

Lampert *et al.* (2009) applies the concept of attribute learning to a particular scenario in object class recognition, where the set of object classes seen during training and test are guaranteed to be disjoint, and has hence been called *zero-shot* recognition. The approach is based on providing an explicit mapping between object classes and an inventory of descriptive visual attributes. The mapping is provided by human supervision (Osherson *et al.*, 1991), and allows to re-use attribute models learned from a set of known training object classes in the context of previously unseen test object classes. In particular, each attribute model consists of a discriminatively trained classifier, distinguishing between images of object classes where the attribute is active versus those where it is inactive. Attribute classifiers are modeled on the level of entire images as SVMs using six different feature channels (color histograms, SIFT (Lowe, 2004), rgSIFT (van de Sande *et al.*, 2010), PHOG (Bosch *et al.*, 2007), SURF (Bay *et al.*, 2008), and local self-similarity histograms (Shechtman and Irani, 2007)) and a sum of corresponding χ^2-kernels. Lampert *et al.* (2009) further propose two distinct probabilistic models for combining individual classifiers to form object class models of unseen test classes, direct attribute prediction (DAP) and indirect attribute prediction (IAP). While for DAP, attribute activations are determined from individual attribute classifiers, IAP infers attribute activations indirectly, from classifiers predicting affiliations to object classes observed during training (standard multi-class classification). In both cases, test class affiliations are determined by the deterministic mapping between active attributes and test object classes. In order to evaluate their approach, Lampert *et al.* (2009) propose a novel data set for zero-shot recognition, termed the Animals-with-Attributes (AwA) data set, and demonstrate promising performance of attribute-based zero-shot recognition in comparison to a standard multi-class scheme that uses training images of the test classes of interest.

Farhadi *et al.* (2009) extend attribute-based object class modeling in various directions compared to Lampert *et al.* (2009), advocating a paradigm shift in recognition from "naming" (by object classes) to "describing" (by attributes). On the level of individual attributes, their approach is generally similar to Lampert *et al.* (2009) in that it also uses attribute classifiers that can potentially draw from a large pool of different features. In contrast to Lampert *et al.* (2009), Farhadi *et al.* (2009) explicitly decouple attribute classifiers from incidentally correlated visual properties by feature selection, and extend the limited pool of semantically meaningful attribute classifiers by randomly selected object class discriminators. On the level of object classes, they

exploit the attribute model further by reporting unusual (i.e., inactive despite common for an object class) and unexpected (i.e., active despite uncommon) attributes. Performance is evaluated on a newly proposed data set, which is based on PASCAL VOC 2008, and provides attribute annotations for all annotated object bounding boxes. For object class recognition, the attribute model is shown to significantly reduce the required amount of training images for a given performance level. Zero-shot recognition is phrased as a nearest neighbor problem between pre-specified object class-attribute associations.

Wang and Forsyth (2009) propose a joint framework of object classes and attributes, based on co-trained multi-instance learners. Training proceeds in a weakly supervised fashion, from image-level object class and attribute labels. It exploits the required agreement between trained object class and attribute detectors with respect to image localization as a search space reduction, and additionally uses top-down saliency and bottom-up homogeneity as cues that guide the search. Experiments are conducted on a novel data set providing object class-attribute pair labels, such as "red car" or "yellow dress", and includes the transfer of learned models to previously unseen object class-attribute combinations.

Kumar *et al.* (2009) apply attribute models to face verification, i.e., determining whether two images depict the face of the same person, and related to object instance recognition rather than object class recognition. Similar in spirit to learning randomized discriminators (Farhadi *et al.*, 2009), an inventory of semantically meaningful attribute classifiers (such as gender, age, or skin color) is enhanced by introducing simile classifiers, which determine whether certain aspects of a face image are similar to the corresponding aspects of a reference person. The concrete implementation of attribute and simile classifiers is again based on combining a multitude of different feature channels with powerful learning machines (SVMs).

Exploiting object class hierarchy. Another line of research employs knowledge about the structure of the space of object classes in order to facilitate recognition. The structure is given as a hierarchy of varying depth among classes, hinting on commonalities between classes that share a common ancestor. Following this basic idea, Levi *et al.* (2004) propose to share features between object classes of more general categories, e.g., between apples and oranges, which are both fruits. Feature sharing is thereby guided by category affiliation, and not entirely data-driven (Torralba *et al.*, 2004). The selection of each individual feature is determined by a category-specific estimate of the expected error for that feature, implemented as a modified weak classifier selection rule in AdaBoost (Freund and Schapire, 1997). Experiments are conducted on a novel data set of object classes of two general categories, fruits and indoor office objects. The reported results suggest that using category-specific error estimates improves recognition performance, in particular if few training instances are available.

Zweig and Weinshall (2007) associate individual classifiers (Bar-Hillel and Weinshall, 2008) to each node in a manually constructed hierarchy of object classes, corresponding to different re-samplings of the training data. Improved recognition

performance is confirmed both theoretically and experimentally for a stacking-based combination of classifiers of different hierarchy levels. In particular, the error probability of the combined classifiers is shown to be smaller than the mean error probability of the constituent classifiers. Zweig and Weinshall (2007) further conduct extensive experiments that highlight the different qualities of classifiers from different hierarchy levels. First, while high level models are superior in terms of recall, low level models typically achieve higher precision. Second, for scarce training data, combining a low level classifier with either a parent or sibling classifier proves beneficial. Third, combined models almost always perform better than either constituents.

Marszalek and Schmid (2007) use WordNet (Fellbaum, 1998) to construct an object class hierarchy, by first querying WordNet for the object class labels of interest, and traversing the graph starting from the identified synsets via hypernymy ("is a" relations) and meronymy ("part of" relation) links. The resulting WordNet subgraph is further pruned by removing all nodes not reachable from a minimal hyponym with respect to all object classes of interest, effectively restricting intermediate nodes to a limited domain. While, as in Zweig and Weinshall (2007), classifiers are associated to the nodes in the graph, the training data for each classifier is defined differently. Training classifiers in a differential fashion leads to the notion of conditional classifiers, which can distinguish among several hyponyms of one and the same hypernym, assuming that the hypernym has already been verified. For recognition, conditional classifiers are evaluated in succession. The given experimental evaluation consists of two parts. In the first part, the semantic hierarchy classifiers are shown to outperform a standard multi-class baseline by a small margin on the PASCAL VOC 2006 data set. In the second part, Marszalek and Schmid (2007) report improved generalization of semantic hierarchy classifiers to non-leaf object classes over baseline methods, and conclude with an initial attempt at zero-shot recognition.

Bayesian priors. The Bayesian paradigm of combining prior knowledge with knowledge arising from observed data to yield posterior estimates gives rise to a specific flavor of knowledge transfer. Transferable knowledge is modeled as a prior probability density. Miller *et al.* (2000) apply this approach to handwritten character recognition, by estimating a kernel probability density of image transformations resulting from aligning all training images depicting the same character (they coined the term congealing for this alignment procedure). The density estimate then signifies a probabilistic characterization of the expected variation in the appearance of any given character, and can thus be transfered to previously unseen characters. In particular, Miller *et al.* (2000) generate artificial training data from a single training example of a previously unseen character by applying various transformations represented by the density, and demonstrate superior recognition performance using this prior knowledge.

Fei-Fei *et al.* (2006) encode prior knowledge about object classes as a prior probability density over the parameter space of object class models. The constellation model of Fergus *et al.* (2003) is used as the basis of the approach, since it represents

both object class appearance and shape as parametric densities, for which conjugate prior distributions exist, and hence allow for efficient inference. The prior over model parameters is learned from three diverse exemplar object classes (spotted cats, faces, and airplanes), resulting in a generic characterization of plausible object class models that can be further specialized by observing training examples of specific object classes. Adding training examples yields a posterior estimate over models, which is used for making predictions (i.e., do recognition) by integrating over model parameters. Experiments are conducted mainly on the Caltech-101 data set, evaluating the performance of the proposed Bayesian approach in comparison to standard maximum likelihood (ML) and maximum a posteriori (MAP) variants of the constellation model Fergus *et al.* (2003). The results confirm the effectiveness of incorporating prior knowledge, in particular for scarce training data.

2.2.2 Additional sources of information

Knowledge transfer using additional sources of information beyond purely visual information often relies on language resources, such as textual image captions and tags, geotags, manually or automatically created ontologies, or specific linguistic knowledge bases (e.g., the world wide web). The work presented in Chapter 8 exploits linguistic knowledge bases for determining possible sources and targets of knowledge transfer, and is thus related to works in this field.

Image captions and tags. In world wide web pages and community photo collections, images are often associated caption texts or user-provided tags, containing potentially useful information about the image content. As a consequence, numerous approaches have been proposed that subsume image content and related text in joint models. Barnard *et al.* (2003) examine different variants of probabilistic models, which achieve a coupling between image and text features by assuming an underlying joint generative process. The examined models are based on latent concepts giving rise to both image features and text, and trained using expectation maximization (Dempster *et al.*, 1977). Experimental results are reported on a subset of the Corel image data set, and evaluated with respect to both image- and region-level labeling accuracy.

Similar in spirit, Li *et al.* (2009) propose a joint framework for classification, annotation, and segmentation of images, using associated tags. The proposed model is based on a probabilistic generative process, combining the texture of image segments, the appearance of image patches, and occurrence of tags. In particular, tags can be subject to noise, which is explicitly accounted for in the model by switching variables. The model is learned via collapsed Gibbs sampling, and demonstrated to yield excellent performance in scene classification, image annotation, and image segmentation tasks.

Jamieson *et al.* (2010) propose to learn structured appearance models of objects from unstructured collections of captioned images, based on co-occurrences of visual features and words in caption texts. The approach is targeted towards

instance recognition rather than object class-level recognition, and uses spatially arranged collections of interest point-based local features as a representation of object appearance. Guided by their co-occurrence with words in caption texts, groups of local features are iteratively grown, until stable object models have been formed. The proposed method is employed to auto-annotate images of architectural landmarks and sports events.

Kalogerakis *et al.* (2009) consider the problem of geolocating sequences of non-geotagged images, taken by a single user in the course of traveling. The proposed method is based on the combination of image evidence and a prior distribution on plausible human travel routes, learned from a collection of geotagged Flickr images.

Ontologies. In contrast to unstructured textual resources, such as text corpora, ontologies provide a formal representation of knowledge, giving a concise representation of the relationships between concepts of a given domain. Ontologies thus lend themselves as conveniently accessible sources of semantic information, once they have been built. Popescu *et al.* (2007) use an ontology automatically built from WordNet (Miller, 1995) to assist both text- and content-based image retrieval. Specifically, the ontology is used to prune the search space for retrieval, in order to improve computational efficiency.

Wang *et al.* (2008) focus on the automatic construction of ontologies to aid web image retrieval, and propose an inferencing mechanism based on spreading activation theory for relevance ranking. Multi-modal ontologies are constructed by following taxonomy relations in Wikipedia articles, and linking concepts to visual classifiers. Performance is evaluated by re-ranking Google image search results.

Linguistic knowledge bases. In the field of image retrieval and scene classification, researchers have started using textual information beyond image captions and tags. Delezoide *et al.* (2008) use world wide web search engines (Exalead) and Flickr tags to mine co-occurrence statistics of object and scene classes. In particular, their approach estimates the distributions of object classes conditioned on the presence of scene classes, linking individual object class and scene classifiers in a probabilistic framework for contextual recognition. The approach is evaluated on a novel data set of animal object classes in the context of their natural surroundings, and shown to outperform models not contextual information.

Boiy *et al.* (2008) aim at determining the visualness of linguistic entities from text corpora, i.e., identify both nouns that refer to visual objects or persons and adjectives that refer to a physical and visual qualification of an object or person. The suggested approach is purely language-driven, by contrasting two different corpora, one of which is assumed to contain mostly visual entities, while the other does not, based on statistical association techniques. WordNet is used as a complementary source of information, propagating visualness along hierarchy paths from annotated seed synsets, and compared to the corpus-based approach. Both approaches are shown to yield satisfactory performance in classifying linguistic entities according to their expected visualness.

While the focus of Barnard and Yanai (2006) is the joint modeling of images and associated tags, they additionally propose a method to determining the visualness of linguistic concepts, similar to Boiy *et al.* (2008). The method uses Google image search to obtain an initial set of potentially relevant images for a given concept. Assuming a generative process linking the concept and image region appearance, visualness can be determined by measuring the entropy of their joint distribution. Low entropy is interpreted as a hint towards a visual concept.

2.2.3 Generalization beyond basic-level categories

The generalization of object classes beyond basic-level categories (Rosch *et al.*, 1976) is often motivated by the large number of these categories (tens of thousands), which can be avoided by generalization. Certain functions, e.g., are supported by a large variety of objects belonging to different basic-level categories. As a consequence, all these objects can be subsumed in a single, functional characterization, rather than referring to each individual basic-level category. In Chapter 4, we present an approach for object class recognition according to functional classes, and hence list related work in this section.

Functional categories. The seminal work of Stark and Bowyer (1991) designs a system for classifying face-vertex geometry descriptions of 3D objects according to functional categories, formulated as a constraint satisfaction problem. Functional categories are defined by hierarchical graphs (category definition trees), where nodes represent sub-functions, annotated with numerical constraint values, and edges represent dependencies between sub-functions. Classifying a given geometry description amounts to first finding possible realizations of all sub-functions in the list of faces, and then verifying the specified sub-function dependencies. The connection between functional description and geometry is established by measurable procedural knowledge primitives (PKPs), such as relative orientation, dimensions, stability, proximity, and clearance. The system is tested on a database of manually designed objects, for the exemplary functional category chair. The corresponding category definition tree has been manually designed. Stark *et al.* (1993) transfer the concepts developed by Stark and Bowyer (1991) to laser range images, acquired by a mobile robotic platform. Input geometry is represented as "object plus unseen space" (OPUS), extending the previous work by explicitly modeling the uncertainty about occluded object portions. As a consequence, reasoning about functional categories comprises three potential assessments, functional, possibly functional, and not functional, and can additionally suggest useful viewpoints in the spirit of active vision. The proposed approach is evaluated on a collection of range images of manually assembled wooden object models.

Building on prior work in functional categorization (Stark *et al.*, 1993), Green *et al.* (1995) propose an approach for categorizing articulated objects, represented as sequences of polyhedral boundary descriptions. Categorization is phrased as a two-step procedure, in which an articulated shape model is first hypothesized from a

sequence of exemplary boundary descriptions, demonstrating possible articulations, and then examined with respect to supported functional properties. The approach is demonstrated on a collection of manually designed instances of the functional category scissors.

Rivlin *et al.* (1995) build upon the ideas of Stark and Bowyer (1991) with respect to functional categorization, but follow a part-based approach for recognition (Pentland, 1986), based on the recovery of volumetric shape primitives (superquadrics). As a result, the proposed system is applicable to both laser range and intensity image data, and offers limited robustness to background clutter and partial occlusion. Objects are modeled on two different levels, shape and function. On the shape level, an object consists of a spatially constrained arrangement of shape primitives. On the function level, an object is composed of multiple functional primitives and their relations. The mapping between shape and function is defined on the level of individual primitives and relations, and assumed to be many-to-one from shape to functional primitives. Recognition proceeds by extracting volumetric primitives and their spatial relations from an input image (Dickinson *et al.*, 1992), inferring corresponding functional primitives and relations, and matching the inferred functional description to a database of descriptions (unexpected object recognition). Prior knowledge about the functional category of interest can be used to prune the search space of local primitive recovery (expected object recognition). Qualitative experimental results are reported on intensity images depicting manually designed tools in lightly cluttered scenes.

While prior work has mostly focused on visual sensory input, Bogoni and Bajcsy (1995) adds force-torque sensors and manipulation capabilities to the exploration of functional categories, following an active perception paradigm. In particular, functional categorization is based on observing interactions between a robot manipulator and an object of interest, formalized by means of discrete event system theory (DES). As in prior work, visual sensory input is pre-processed to yield a superellipse primitive-based object representation. The proposed combination of visual and force-torque sensors is demonstrated using the example of piercing actions, where a target object is penetrated by a tool.

More recently, Kjellström *et al.* (2008) propose the simultaneous recognition of manipulation actions and manipulated objects, for the mutual benefit of both. Actions are defined by shape descriptions of manipulated object and manipulator (a human hand), and their relative positions, both measured over time. Action as well as object recognition is performed using a joint probabilistic model, trained in a discriminative fashion, from labeled video sequences. The correlation between actions and objects is captured in a hierarchical conditional random field (CRF) (Lafferty *et al.*, 2001) model, and experimentally shown to be beneficial for the recognition of both.

Robotic grasping. Saxena *et al.* (2007) consider the problem of robotic grasping of previously unseen objects and object classes. The proposed approach is based on learning models for 2D image projections of good grasping points in a supervised fashion, and using these models in connection with multiple cameras to obtain a

sparse triangulation of potential grasping points in 3D, circumventing the need for an expensive and error-prone dense reconstruction. Models are learned using synthetic object models with manually provided grasping point labels, in combination with logistic regression classifiers. Grasping point predictions of individual views are fused in a probabilistic framework, using maximum a posteriori (MAP) inference. The reported experimental evaluation comprises both a quantitative analysis of potential grasping success on synthetic images and real grasping experiments using robot manipulators of varying degrees of freedom. The experiments suggest the successful application of learned grasping point models (from kitchen objects) on previously unseen object classes (office objects).

2.2.4 Relation to own work

In this section, we highlight the relations between the work presented in this thesis and related work, for each individual chapter, as concerns knowledge transfer in object class recognition. The corresponding relations for general object class recognition are subject to Section 2.1.6.

Chapter 3: Local features for classes of geometric objects. Most prior work in visual knowledge transfer uses a variety of different feature channels in combination with learning algorithms that select relevant channels (Lampert *et al.*, 2009; Farhadi *et al.*, 2009; Kumar *et al.*, 2009). While some approaches include features that can be considered remotely shape-based (Bosch *et al.*, 2007), none of them explores features based on discrete contour segments, such as k-AS (Ferrari *et al.*, 2008). This is in contrast to our evaluation of local features presented in Chapter 3, which explicitly focuses on shape-based features, including k-AS. These constitute the basis for our spline-based representation of transferable part shape (Chapter 5).

Chapter 4: Functional object class detection. Knowledge transfer in terms of generalization across basic level categories has been sparsily explored in the literature, mostly in connection with functional categorization or object affordances (Winston *et al.*, 1983; Stark and Bowyer, 1991; Stark *et al.*, 1993; Green *et al.*, 1995; Rivlin *et al.*, 1995). While the work presented in Chapter 4 is fueled by a similar motivation, it differs largely in the quality of the image data used for recognition experiments (simplistic scenes of manually assembled objects vs. real world images of a standard benchmark data set (Ferrari *et al.*, 2006b)). This is in correspondence with the different starting points assumed for prior work and Chapter 4, respectively. Prior work typically focuses on powerful representations of complex functions, requiring equally complex shape representations which have proven difficult to extract from realistic image data. The work presented in Chapter 4, on the other hand, starts from a robust local shape feature-based object class detector, and explores its use in functional categorization, at the cost of less powerful modeling of functions. The focus lies on functional classes related to grasping, which has been an actively pursued direction in robotics literature (Saxena *et al.*, 2007). Also in contrast in

prior work, we suggest to learn functional categories by observing few prototypical human-object interactions, rather than maintaining an explicit physical model of possible functions (Bogoni and Bajcsy, 1995). Kjellström *et al.* (2008) follow a slightly different direction, by classifying objects and manipulative actions simultaneously, requiring the simultaneous observation of both.

Chapter 5: Shape-based object class model for knowledge transfer. The object class model presented in Chapter 5 is specifically tailored towards knowledge transfer, by choosing a probabilistic formulation that factors into separate components for local part shape, spatial layout, and symmetry relations. The specific form of the chosen component densities allows for an incremental inclusion of transfered knowledge in the spirit of Bayesian priors, as suggested by Miller *et al.* (2000); Fei-Fei *et al.* (2006). In contrast to prior work, the factorized nature of the model allows to transfer knowledge on different levels, either completely, or restricted to proper subsets of components. Equally, transferable components can be addressed explicitly, and flexibly combined to yield new combinations. In the current implementation, the resulting increased flexibility comes at the cost of having to specify manually which components should be transfered between object classes, while prior work allows to infer transferable properties by joint learning (Ben-David and Schuller, 2003; Torralba *et al.*, 2004; Amit *et al.*, 2007) or drawing from a large number of object classes known beforehand (Fink, 2004; Bart and Ullman, 2005b,a). A similar argument applies to required part-level annotations of training images (Sudderth *et al.*, 2008).

Chapter 6: Shading cues for object class detection. Despite the model-based shading cues presented in Chapter 6 being inherently generic and thus potentially transferable in nature, we have not yet confirmed this intuition experimentally. In contrast to this, Nillius *et al.* (2008) demonstrate the applicability of model-based shading cues to a limited collection of different material and objects, such as granite pillars, wooden trunks, and marble spheres. Re-using generic visual cues as building blocks in a variety of more specific models is also similar in spirit to early works based on shape primitives (Nevatia and Binford, 1977; Pentland, 1986; Dickinson *et al.*, 1992), although limited to simplistic objects and scenes.

Chapter 7: Learning shape models from 3D CAD data. The work presented in Chapter 7 builds upon ideas from the early days of computer vision, by making a direct connection between three-dimensional object models and real world images (Nevatia and Binford, 1977; Marr and Nishihara, 1978; Brooks *et al.*, 1979; Pentland, 1986; Lowe, 1987). In contrast to more recent work using 3D models (Liebelt *et al.*, 2008), it circumvents the costly combination of photorealistic rendering and subsequent learning of relevant image gradients by choosing a shape-based abstraction of object appearance that is shared between 3D models and real world images. Using this abstraction, we show improved multi-view recognition performance in comparison to state-of-the-art (Gill and Levine, 2009; Su *et al.*, 2009; Liebelt and

Schmid, 2010) on the 3D object classes data set (Savarese and Fei-Fei, 2007). Specifically, our model, being trained entirely from 3D CAD models of the object class of interest, outperforms prior work that either combines geometric information from 3D models with appearance information from real world training images (Liebelt and Schmid, 2010), or which is solely based on real world training images (Gill and Levine, 2009; Su *et al.*, 2009). We attribute the superior performance of our model to the tight coupling between geometric layout and discriminative part shape (since both are trained from the same 3D CAD models), which circumvents the need for establishing correspondences between different viewpoints (Leibe *et al.*, 2006a; Savarese and Fei-Fei, 2007, 2008; Sun *et al.*, 2009; Su *et al.*, 2009; Arie-Nachimson and Basri, 2009). In contrast to Liebelt and Schmid (2010), where object parts are defined by means of a regular grid structure, our part representation is currently based on semantic part-level annotations of 3D CAD models.

Chapter 8: Semantic relatedness for knowledge transfer. Knowledge transfer has been recognized as an important tool for scalable recognition. The work presented in Chapter 8 builds upon two different approaches proposed for knowledge transfer, namely, attribute-based object class models (Ferrari and Zisserman, 2007; Lampert *et al.*, 2009; Farhadi *et al.*, 2009; Kumar *et al.*, 2009), and object class models based on inter-class distances (Fink, 2004; Bart and Ullman, 2005b). In particular, it uses the direct attribute prediction (DAP) model of Lampert *et al.* (2009) to relate object classes and descriptive attributes, and proposes a variant of the DAP formulation applicable to inter-object class distances. In contrast to prior work, associations between object classes and attributes are not manually specified (Osherson *et al.*, 1991), but determined fully automatically, using semantic relatedness measures in connection with linguistic knowledge bases. Furthermore, we extend the distance-based object class model for the challenging task of zero-shot recognition, which we evaluate on the publicly available Animals-with-Attributes (AwA) data set (Lampert *et al.*, 2009). As concerns semantic relatedness, we investigate the use of various different linguistic knowledge bases combined with state-of-the-art measures proposed by the NLP community (Lin, 1998; Berland and Charniak, 1999; Kilgarriff and Grefenstette, 2003; Budanitsky and Hirst, 2006; Gabrilovich and Markovitch, 2007; Zesch and Gurevych, 2010), which goes beyond the often limited use of language resources in the computer vision literature (Popescu *et al.*, 2007; Li *et al.*, 2009; Jamieson *et al.*, 2010).

3 LOCAL FEATURES FOR CLASSES OF GEOMETRIC OBJECTS

Contents

VISUAL FEATURE REPRESENTATIONS constitute the basis of all object class recognition systems, providing the necessary abstraction for making image information accessible through machine vision algorithms. In the context of knowledge transfer, shape features appear to be particularly important, since the shape of an object or an object part often constitutes a more generic, and thus more likely transferable representation than appearance, as we argue in the introductory example of Section 1.1. This chapter thus provides an extensive evaluation of current state-of-the-art local feature detectors and descriptors, focusing on shape-based representations. Design decisions concerning the choice of local feature representations in later chapters (chapters 4 and 5) are based on the results of this evaluation.

3.1 INTRODUCTION

Historically, the recognition of geometric objects such as cups and tables has been an important focus of object recognition (Mundy, 2006). In recent work however, the recognition of geometric objects is largely underrepresented with some notable exceptions (Ferrari *et al.*, 2006a; Opelt *et al.*, 2006). Many recent and successful object

Figure 3.1: Example images from the *Shape* ((a), (b)), *Shape2* (c), and *Caltech-256* (d) data sets.

recognition and categorization approaches are based on local *appearance* features (Fergus *et al.*, 2003; Csurka *et al.*, 2004; Mikolajczyk *et al.*, 2006). These approaches tend to perform well when enough local appearance information can be found, as, e.g., for cars and motorbikes. Incorporating spatial information about feature distributions often proves helpful, even though surprisingly good results have been reported for simple bag-of-words approaches that neglect feature location altogether.

The main question we address in this chapter is how these local feature approaches transfer to the recognition of more *geometric* objects. We therefore compare the statistics of various successful appearance features (such as SIFT (Lowe, 2004) and GLOH (Mikolajczyk and Schmid, 2005)) with features that are more geometric in nature (such as Geometric Blur (GB) (Berg and Malik, 2001) and Shape Context (SC) (Belongie *et al.*, 2000)). We also include the recently proposed k-Adjacent Segments (k-AS) (Ferrari *et al.*, 2006a) which have been designed to explicitly code information about the geometric layout of an object. The evaluation is performed on two complementary data sets: a new data set containing over 700 images of 10 geometric object classes, and subsets of *Caltech-256* (Fei-Fei *et al.*, 2004) containing object classes with important local appearance statistics. One of the surprising results of this evaluation is that performance differences between features on these two rather distinct data sets are less pronounced than one might expect.

Secondly, we analyze the performance of these features in a general object classification setting based on local features, again using the two distinct databases of object classes. As a first baseline we use a Naïve Bayes classifier, where individual features contribute independently to the classification result. As spatial information might be particularly important for the recognition of geometric objects, we use a second baseline, the so called localized bag-of-words representation. It allows to gradually add location information to the object representation, and quantify the

contribution of location to classification accuracy.

The results of the first baseline experiments are mostly consistent with what is suggested by the feature statistics evaluation. The introduction of (weak) spatial information however results in a more significant performance boost than caused by differences between individual features alone. This performance boost notably differs between features. Without any spatial information, SIFT and GLOH in combination with the Hessian-Laplace interest point detector perform best on average. When spatial information is added, geometric features, namely Geometric Blur and *k*-Adjacent Segments, can outperform all other features.

The first contribution of this chapter is the introduction of a novel data set of 10 geometric object classes. The second is the evaluation of different appearance as well as geometric features on two distinct data sets. Third, we compare two clustering schemes which have been used in the literature by some of the most successful recognition approaches. Fourth, we give results for two baseline methods for local feature-based recognition with and without spatial information.

3.2 RELATED WORK

Local appearance-based features have received a lot of attention in the literature. Many general feature evaluations exist, typically focusing on criteria based on correspondence matching (Mikolajczyk and Schmid, 2005; Mikolajczyk *et al.*, 2005b). Including the notable exception of (Mikolajczyk *et al.*, 2005a), comparably little work has been done on feature evaluation in the context of object class recognition.

Similarly, the explicit treatment of shape-based local features is non-sufficient. The Shape Context descriptor has been studied in the context of pedestrian detection (Seemann *et al.*, 2005). Moreels and Perona (2005) compares its performance to SIFT, PCA-SIFT (Ke and Sukthankar, 2004), Differential Invariants (Schmid and Mohr, 1997) and Steerable Filters in a 3D matching framework. Leibe and Schiele (2003) builds upon global image representations, and compares the performance of texture- and contour-cues in a multi-class classification task. Apart from Ferrari *et al.* (2006a), we are not aware of any evaluation including *k*-Adjacent Segments.

We build upon the work of Mikolajczyk *et al.* (2005a), and base our evaluation on a mid-level clustering representation used by many local feature approaches in object recognition. In fact, we reproduce their major findings by evaluating the best performing features on a subset of *Caltech-256*. Our evaluation goes beyond their work w.r.t. the following aspects: 1) our evaluation explicitly includes shape-based features such as Geometric Blur, Shape Context and *k*-Adjacent Segments; 2) we do experiments over two complementary data sets of shape- and appearance-based classes (notably introducing the shape-based data-set), and 3) we report results over two general baseline recognition frameworks with and without spatial information.

3.3 DATA SETS

We shortly present the data sets used in our experiments.

Shape. We introduce a novel collection of 724 images showing single objects of 10 *geometric* object classes, i.e., objects for which shape and geometric layout of object parts determine class affiliation rather than local appearance. Figures 3.1 (a) and (b) show two examples of each of the classes *cup*, *fork*, *hammer*, *knife*, *mug*, *pan*, *pliers*, *pot*, *saucepan*, and *scissors*. All images are roughly aligned w.r.t. position and viewpoint. The data set exhibits high intra-class variability, and is challenging for recognition. In order to provide a platform for best-case evaluations, objects are recorded in front of a clean, white background. For experiments, we randomly pick 20 images from each class as a training set, and 10 images of each class for testing.

Shape2. *Shape2* contains 100 additional images to provide more realistic test data for the above classes (see Figure 3.1 (c)). While the images contain a single object each the background and image quality vary greatly. Both *Shape* and *Shape2* data sets can be downloaded from our web-page[3].

Caltech-256. We primarily use a 10 class subset of *Caltech-256* (Fei-Fei *et al.*, 2004) that we find is characterized by local appearance statistics, namely *accordion*, *crab*, *cannon*, *electric guitar*, *euphonium*, *gramophone*, *inline skate*, *revolver*, *watch*, and *windsor chair* (see Figure 3.1 (d)). We further report results for a larger pool of object classes provided by *Caltech-256*, but restrict ourselves to two random subsets of 20 and 40 classes, respectively, for computational reasons.

3.4 LOCAL FEATURES

We briefly introduce the features and interest point detectors used in our comparison. The shape related features are *k*-Adjacent Segments (Ferrari *et al.*, 2006a), Geometric Blur (Berg and Malik, 2001), and Shape Context (Belongie *et al.*, 2000); the appearance-based region descriptors are SIFT (Lowe, 2004) and GLOH (Mikolajczyk and Schmid, 2005). As interest point detectors, we employ Harris-Laplace (Mikolajczyk and Schmid, 2004), Hessian-Laplace (Mikolajczyk *et al.*, 2005b), and Salient Regions (Kadir *et al.*, 2004).

3.4.1 *k*-Adjacent Segments (*k*-AS)

k-Adjacent Segments have been proposed as an extension to *contour segment networks*, a graph-based method for template-matching hand-drawings to image databases (Ferrari *et al.*, 2006b). Ferrari *et al.* (2006a) demonstrates how *k*-AS can be incorporated

[3] http://www.mis.informatik.tu-darmstadt.de

into a general object recognition framework.

We extract k-AS features from an image using the original implementation[4]. First, *edgels* are detected via the Berkeley natural boundary detector (Martin *et al.*, 2004). Second, neighboring edgels are chained, and further linked to form L-, T- and higher-order junctions. Last, edgel-chains are replaced by straight line approximations (*contour segments*), and joined into a global *contour segment network* for the image.

A k-AS descriptor d then describes the geometric layout of a group of k adjacent segments in that network. For each such group, one segment is picked as reference, and the layout of the others described relative to that reference segment. In particular, the descriptor encodes relative segment positions p_i, orientations o_i and lengths l_i. Let segment #1 be the reference segment, then the descriptor is

$$d = (p_2/N, \ldots, p_k/N, o_1, \ldots, o_k, l_1/N, \ldots, l_k/N), \tag{3.1}$$

where N is a normalization factor that renders d invariant to scale. We make descriptors invariant to in-plane rotation by rotating them around $(0, 0)$ by o_1, and excluding o_1 from the descriptor. The dimensionality of k-AS features is $n = 4 * k - 2$. For the typical choices of $k \in \{2, 3\}$, n is 6 or 10, rendering k-AS a comparably low dimensional descriptor.

k-AS features differ from others considered in this chapter in several ways. Notably, they do not match the standard scheme of describing local image regions around interest points. Instead, they utilize shape information from the whole image, and thus have the potential to capture the characteristic geometric layout of an object at the cost of sacrificing local appearance information. Although k-AS mostly represent local groups of contour segments, we also observed that descriptors related distant contour segments on opposing object boundaries.

3.4.2 Local region descriptors

We briefly present the local region descriptors used in our experiments. Local region descriptors encode information about image patches centered at interest points.

Geometric Blur (*GB*). We use the original implementation[5] of the Geometric Blur (Berg and Malik, 2001) region descriptor from Berg *et al.* (2005). Geometric Blur first extracts $c = 4$ channels of oriented edge energy (Morrone and Burr, 1988) to obtain a sparse signal S. In S, the region centered at interest point location x_0 is blurred with a spatially-varying Gaussian kernel G_d to obtain the Geometric Blur $B_{x_0}(x) = S * G_{\alpha x+\beta}(x_0 - x)$. $B_{x_0}(x)$ is then sub-sampled over all channels at n distinct locations in a circular grid. The final descriptor is the concatenation of all $c \times n$ samples. Throughout all experiments, we use the standard values for $\alpha = 0.5$, $\beta = 1$ and $n = 51$, resulting in a descriptor of length 204.

[4] http://www.vision.ee.ethz.ch/~ferrari/release-kas.tgz

[5] http://www.cs.berkeley.edu/~aberg/demos/gb_demo.tar.gz

Shape Context (*SC*). Shape Context (Belongie *et al.*, 2000) is originally based on edge information. For a given interest point location, it accumulates the relative locations of nearby edge points in a coarse log-polar histogram. We use the implementation[6] of Mikolajczyk and Schmid (2005), and compute a histogram containing 9 spatial bins over 4 edge orientation channels. Bin size increases w.r.t. distance from the interest point center. Note that this is similar in spirit to spatially varying blur, but results in a smaller descriptor (length 36).

SIFT. The Scale Invariant Feature Transform (Lowe, 2004) descriptor is a 3D histogram over local gradient locations and orientations, weighted by gradient magnitude. It uses 4×4 location and 8 orientation bins, i.e., 128 in total.

GLOH. Gradient Location Orientation Histograms (Mikolajczyk and Schmid, 2005) is an extension of the SIFT descriptor. It uses 17 bins for location and 16 bins for orientation in a histogram over a log-polar location grid, and reduces descriptor dimensionality to 128 by PCA. We use the implementation[7] of (Mikolajczyk and Schmid, 2005) for SC, SIFT and GLOH. All descriptors are made invariant to in-plane rotation by aligning the region to the dominant gradient direction before descriptor computation.

3.4.3 Interest point detectors

We compute local region descriptors based on detections of the following interest point detectors: **Harris-Laplace** *(HarLap)* (Mikolajczyk *et al.*, 2005b) is an extension to Harris corners (Harris and Stephens, 1988). It selects corners at locations where a Laplacian attains an extremum in scale-space. The **Hessian-Laplace** *(HesLap)* (Mikolajczyk *et al.*, 2005b) detector responds to blob-like structures. It searches for local maxima of the Hessian determinant, and selects a characteristic scale via the Laplacian as for Harris-Laplace. The **Salient Regions** *(SalReg)* detector (Kadir *et al.*, 2004) identifies local image regions that are non-predictable across scales by measuring entropy over local intensity histograms. We use publicly available implementations for HarLap/HesLap[8], and SalReg[9].

3.5 FEATURE EVALUATION

We evaluate and compare the combined performance of feature detectors and descriptors at three different levels. First, we compute statistics over clusterings of local feature descriptors (codebooks), using two different clustering techniques. Second, we represent objects by means of occurrence statistics over codebook matches, and

[6] http://www.robots.ox.ac.uk/ vgg/research/affine/
[7] http://www.robots.ox.ac.uk/~vgg/research/affine/descriptors.html
[8] http://www.robots.ox.ac.uk/~vgg/research/affine/detectors.html
[9] http://www.robots.ox.ac.uk/~timork/salscale.html

analyze classification performance in a Bayesian framework. Third, we investigate the impact of gradually adding location information to that object representation, by jointly boosting localized histograms of codebook matches over all object categories.

3.5.1 Cluster statistics

We follow the argumentation of Mikolajczyk *et al.* (2005a) and base our evaluation on a mid-level representation of image features common to many computer vision techniques. In particular, we analyze the statistics of clusterings of feature descriptors. A clustering over a set of feature descriptors of a given type is determined by 1) the choice of clustering algorithm, and 2) the choice of a (dis-) similarity measure. We use K-Means as a widely accepted method for 1), and add Reciprocal Nearest Neighbor (RNN) clustering for comparability with Mikolajczyk *et al.* (2005a). RNN is a centroid-based implementation of Hierarchical Agglomerative clustering (Leibe *et al.*, 2006b). For 2), we consistently use Euclidean distance. We are conscious that various clustering techniques and (dis-) similarity measures have been proposed tailored towards specific feature types. We resort to standard ones for the sake of comparability.

Cluster precision. In order to quantify how well a clustering of feature descriptors reflects the separation of object classes, we introduce a refinement to *cluster precision* (Mikolajczyk *et al.*, 2005a). Intuitively, we want to measure to what extent features of a given class a are grouped together by clustering. Original cluster precision therefore memorizes, for each cluster j in which class a dominates, the fraction p_{j_a} of features of class a, and averages these fractions by the number of clusters M dominated by a, i.e., $P_{Ca} = \frac{1}{M}\sum_{j=1}^{M} p_{j_a}$. In our experiments, we found that many clusters with high scores according to p_{j_a} often contained features from only a single object instance. Because we want to give higher scores to feature descriptors that generalize across multiple instances of an object class, we discount such clusters by summing over the fractions of *objects* of class a in cluster j instead of individual features, and weight these fractions by cluster sizes. We obtain

$$P'_{Ca} = (\sum_{j=1}^{N} s_j)^{-1} \sum_{j=1}^{N} s_j p_{j_a}, \tag{3.2}$$

where j now ranges over all N clusters in which *objects* of class a dominate, and s_j is the total number of features in cluster j. High scores (including weights) are obviously obtained by big clusters with features from many instances of a single object class, and low scores by small clusters with few features, but from multiple classes. The combined score for a descriptor is the average over P'_{Ca} over all classes.

3.5.2 Naïve Bayes

The second level of our evaluation builds upon the codebooks we used for measuring cluster precision. It represents objects in terms of occurrence statistics (counts over nearest-neighbor matches) over codebook entries and trains a multi-class-classifier on a training set of such representations. We use an analogous approach to *Multinomial Naïve Bayes* (McCallum and Nigam, 1998) for text classification, and model the posterior distribution of an object class, given occurrence statistics over a codebook, as a multinomial distribution. Let N_{ij} denote the number of occurrences of a feature of class c_j that matches codebook entry w_i. We estimate the likelihood that this codebook entry originates from class c_j as

$$P(w_i|c_j) = \frac{1 + N_{ij}}{W + \sum_{c=1}^{C} N_{ic}}, \tag{3.3}$$

for C different object classes and a codebook of size W, and assuming a simple Laplacian prior.

3.5.3 Localized bag-of-words

While Mikolajczyk *et al.* (2005a) relates the localization properties of a given feature type to entropy over location distributions, we directly measure the impact of adding location information in terms of classification accuracy in a Joint Boosting (Torralba *et al.*, 2004) framework. The object representation on the third level of our evaluation is based on soft-matched histograms of feature occurrences over a codebook, and inspired by Ferrari *et al.* (2006b). We divide a rectangular image region into a grid of cells. For each cell, extracted features are matched to a codebook, and a local histogram over soft-matched codebook entries is computed (the inverse distances between feature and all cluster centroids are used). The representation of an object is the concatenation of all local cell histograms. A Joint Boosting algorithm is trained from object representations of a set of training images using a fixed number of boosting rounds, and tested against an independent set of test images. We use *decision stumps* (Torralba *et al.*, 2004) over histogram bins as weak classifiers for boosting. By varying the number of grid cells g, we regulate the tradeoff between rich feature statistics (small g) and more accurate localization (large g).

3.6 EXPERIMENTAL RESULTS

In the following, we present the results of our experimental evaluation. For the sake of readability, we choose to give separate plots for SIFT, GB, and k-AS, since their comparison is a key contribution of this chapter. That is, in each plot (Figures 3.2, 3.3, and 3.4), we fix the descriptor of interest (SIFT, GB, k-AS), varying the respective detector (respectively varying k for k-AS). Additionally, we give curves for SC and GLOH descriptors that obtain highest scores with varying detectors as part of the

SIFT plots. We also plot non-rotation invariant *k*-AS as a reference, and include the DoG detector proposed for SIFT (Lowe, 2004), using the original implementation[10].

For comparability, all plots in this chapter consistently show results for 10 classes of the respective data set (*Shape*, *Shape2*, *Caltech-256*). For *Caltech-256*, we emphasize the major differences to 20 and 40 class subsets in the text. The complete collection of plots for cluster precision, Naïve Bayes classification, and Localized Bag-of-Words over all data sets, allocating a separate plot for varying descriptors over a single detector, has been published as supplemental material to Stark and Schiele (2007).

3.6.1 Cluster statistics

We measure cluster precision for 9 different compression ratios (*#Features/#Clusters*) ranging from 4 to 20 in steps of 2 over codebooks generated from 200 training images per data set. We give cluster precision plots in Figure 3.2, where each row corresponds to a distinct experiment. The first row (plots (a) to (c)) gives cluster precision results for *Caltech-256*, the second row (plots (d) to (f)) for *Shape*. The third row (plots (g) to (i)) repeats the experiment of row one, using a different clustering algorithm.

As we might intuitively expect , cluster precision generally decreases with increasing compression ratio, i.e., increasing average number of features per cluster. We also note that cluster precision changes substantially if we vary detectors for a given descriptor (high variance within single plots), while it remains relatively stable over varying descriptors for a given detector (lower variance between corresponding curves across different plots).

Caltech-256. We begin by presenting the results for *Caltech-256* (see Figures 3.2 (a) to (c)), and first consider the appearance-based SIFT and GLOH descriptors with varying detectors. We observe that the ordering of detector performance is in fact consistent across all examined descriptors, including GB and SC. HesLap is best, followed by HarLap and SalReg. SIFT and GLOH descriptors perform equally well, and obtain high scores in our comparison. Both obtain highest scores for HesLap. These results are in line with the results reported in Mikolajczyk *et al.* (2005a).

Shape. Surprisingly, these results transfer seamlessly to the *Shape* data set (see Figures 3.2 (d) to (f)). Still, appearance-based SIFT and GLOH obtain high scores, and the ordering of detectors remains the same as for *Caltech-256*. We stress that this stability across data sets is unexpected. We now examine the performance of the shape-based features GB, SC, and *k*-AS. Over both data sets, the precision of GB is slightly higher than that of SIFT for HesLap and SalReg detectors (HesLap-GB is best), and about equal for HarLap. SC is best with HesLap, but slightly worse than SIFT over all detectors. *k*-AS and DoG-SIFT obtain lowest scores. The relative ordering of *k*-AS follows the intuition that discriminative power increases with *k*,

[10] http://www.cs.ubc.ca/~lowe/keypoints/

and decreases with rotation invariance.

Clustering algorithms. Figures 3.2 (g) to (i) correspond to plots (d) to (f), but for Reciprocal Nearest Neighbor clustering. The relative ordering of detectors and descriptors is mainly consistent with K-Means, with the notable exception of GB (Figure 3.2 (h)), where HesLap is inferior to HarLap, and SalReg dominates both for high compression ratios. On average, we obtain higher absolute precisions for K-Means clustering. Reciprocal Nearest Neighbor shows stronger tendencies to yield degenerate clusterings for high compression ratios, where a single large cluster attracts most of the features while leaving others unmatched (singletons). For low compression ratios often used in codebook-based approaches, this is not an issue.

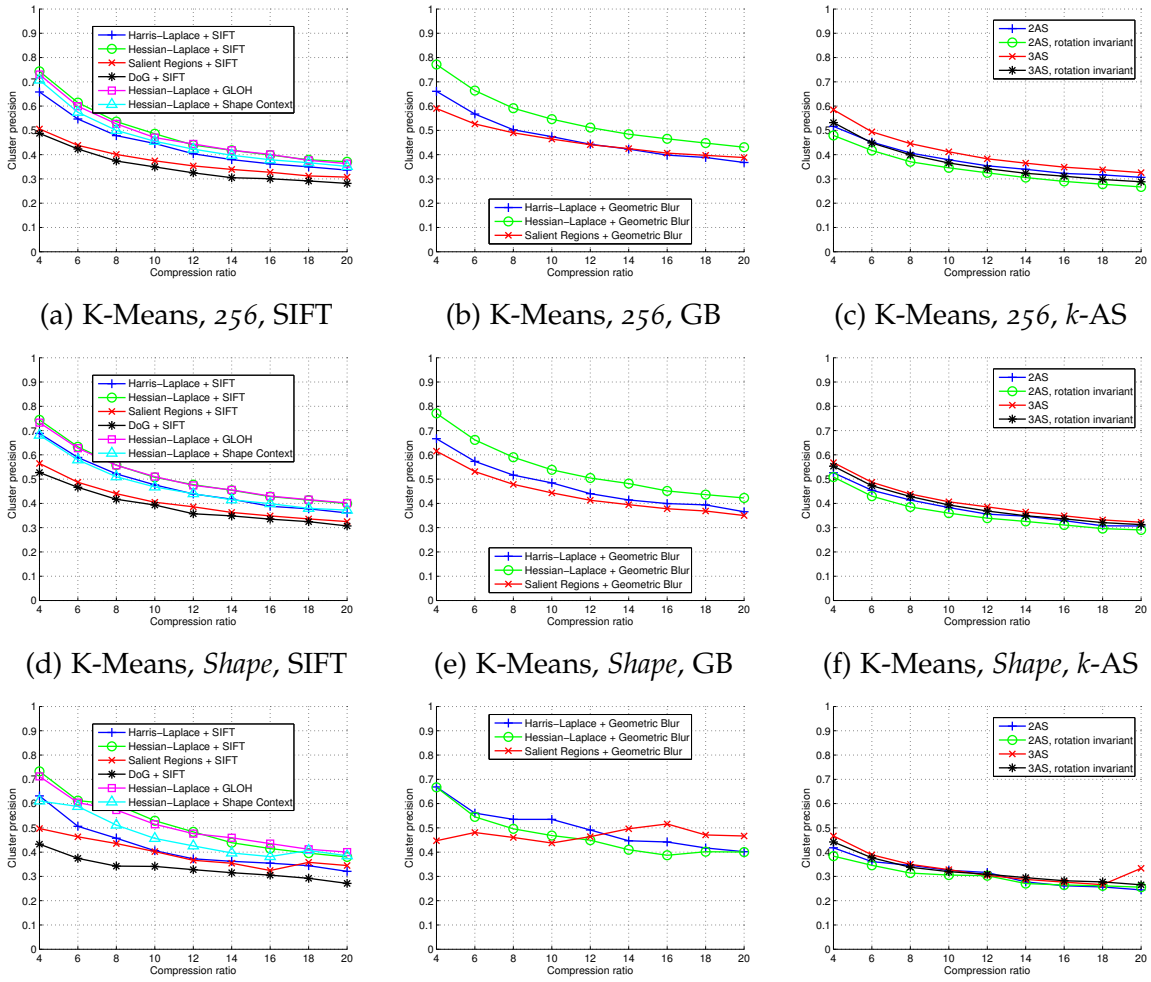
Summary. In summary, appearance- (SIFT, GLOH) and shape-based descriptors (GB, SC) do not show great differences in cluster precision. Both perform comparably well over both appearance-based (*Caltech-256*) and shape-based (*Shape*) data sets. k-AS are worst, but still comparable to DoG-SIFT. These results are fully mirrored on 20 and 40 class subsets of *Caltech-256*.

3.6.2 Naïve Bayes

We measure multi-class-classification accuracy ($\#correctPredictions/\#totalPredictions$) over fixed numbers of clusters from $n = 50$ to 1600, increasing by powers of 2. Dependent on the detector, $n = 1600$ corresponds to compression ratios of 9 (HarLap), 52 (HesLap), 21 (SalReg), 5 (DoG), 5 (2-AS) and 16 (3-AS) for the *Shape* data set. For each feature type and data set, we train a Naïve Bayes classifier on a training set of 200 images (20 per category), and test on an independent test set of 100 images (10 per category). Classifiers are trained on bag-of-word representations (hard-matching against a codebook of size n built from the training set). Figure 3.3 gives the results for Naïve Bayes classification for *Caltech-256* (first row, plots (a) to (c)) and *Shape* (second row, plots (d) to (f)).

Caltech-256. Again, we start with *Caltech-256* (see Figures 3.3 (a) to (c)). The ordering of detectors over all descriptors is mainly consistent with the results for cluster precision: HesLap is typically best, closely followed by HarLap, and SalReg. Overall, the appearance-based feature combination HarLap-SIFT performs best, followed by HesLap-GLOH and HesLap-SC. While GB performs moderately on 10 object classes with HesLap, it performs generally worse than other descriptors for 20 and 40 classes. Rotation invariant k-AS are consistently worse. SC is competitive to SIFT with HesLap, in particular for 20 and 40 classes.

Shape. For the *Shape* data set (see Figures 3.3 (d) to (f)), the order of detectors is consistent with the results for *Caltech-256*; only SalReg gain relative performance in combination with the shape-based GB descriptor. Still, appearance-based HesLap-SIFT and HesLap-GLOH are best. Although GB is comparable with SalReg, it

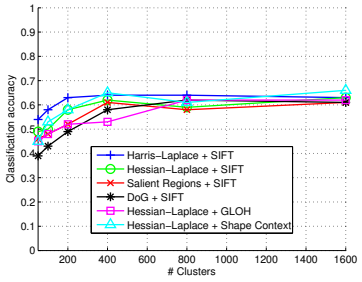


(a) K-Means, *256*, SIFT (b) K-Means, *256*, GB (c) K-Means, *256*, *k*-AS

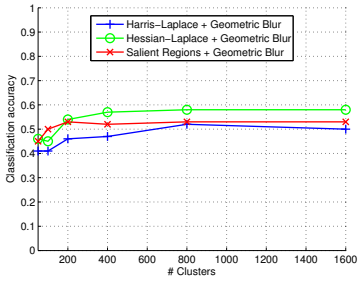
(d) K-Means, *Shape*, SIFT (e) K-Means, *Shape*, GB (f) K-Means, *Shape*, *k*-AS

(g) RNN, *Shape*, SIFT (h) RNN, *Shape*, GB (i) RNN, *Shape*, *k*-AS

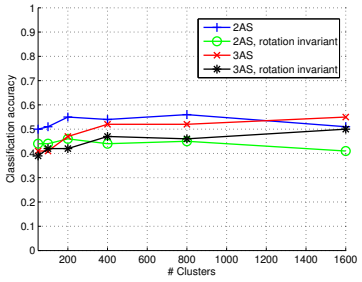
Figure 3.2: Cluster precision for SIFT, GB, and *k*-AS, for *Caltech-256* (first row) and *Shape* (second and third row), using K-Means (first and second row) and RNN clustering (third row), respectively. SIFT-plots include the best GLOH and SC curves.



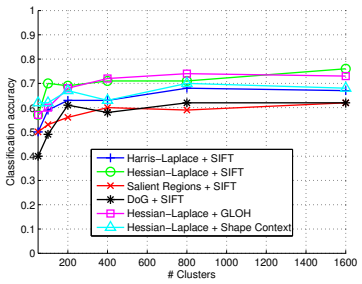
(a) Naïve Bayes, *256*, SIFT



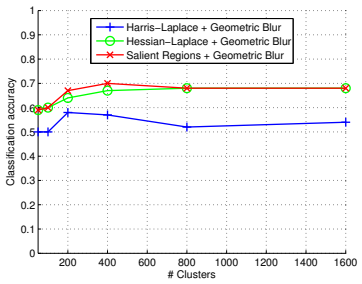
(b) Naïve Bayes, *256*, GB



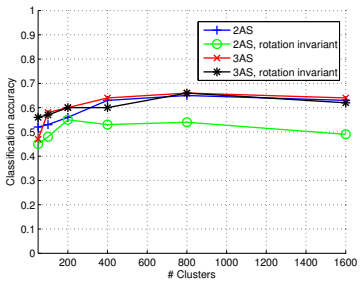
(c) Naïve Bayes, *256*, *k*-AS



(d) Naïve Bayes, *Shape*, SIFT



(e) Naïve Bayes, *Shape*, GB



(f) Naïve Bayes, *Shape*, *k*-AS

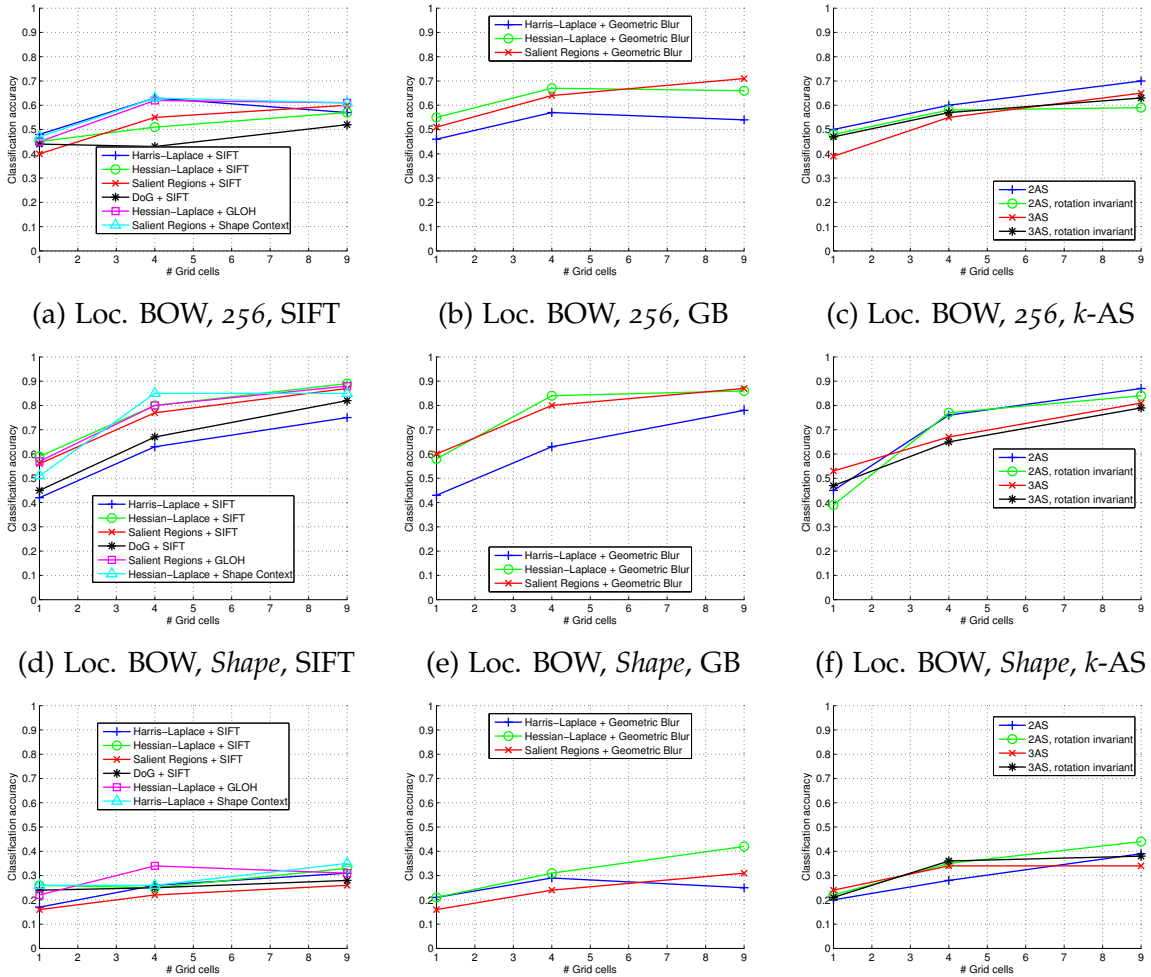
Figure 3.3: Naïve Bayes classification accuracies for SIFT, GB, and *k*-AS, for *Caltech-256* (first row) and *Shape* (second row). SIFT-plots include the best GLOH and SC curves.

performs worse than SIFT with HarLap and HesLap. Notably, it tends to perform better for lower numbers of clusters, which can be explained by its high dimensionality, bearing the risk of over-fitting. SC is generally inferior to SIFT, but superior to rotation invariant *k*-AS. Rotation invariant 2-AS are worst. Rotation invariant 3-AS are slightly better than DoG-SIFT, but can not keep up with SIFT in general.

Summary. To summarize, the differentiation between descriptors with fixed detectors is more pronounced for Naïve Bayes than for cluster precision. In particular, appearance based features lead on average, over both data sets. GB and SC perform on a comparable level to SIFT and GLOH, but only for individual detectors (SalReg for GB, HesLap for SC). *k*-AS exhibit relatively weak discriminative power for Naïve Bayes classification. SC offers a good compromise between strong (GB) and weak (*k*-AS) discrimination.

3.6.3 Localized bag-of-words

We measure classification accuracy as defined for Naïve Bayes for varying numbers of grid cells $g \in \{1, 4, 9\}$, using the same training and test images. We assume known bounding boxes for training *and* test, and use them to anchor histogram grids. We fix the number of clusters to $n = 200$, and obtain histograms of length $g \times n \in \{200, 800, 1800\}$. Figure 3.4 gives the results for Localized Bag-of-Words



(a) Loc. BOW, *256*, SIFT (b) Loc. BOW, *256*, GB (c) Loc. BOW, *256*, *k*-AS

(d) Loc. BOW, *Shape*, SIFT (e) Loc. BOW, *Shape*, GB (f) Loc. BOW, *Shape*, *k*-AS

(g) Loc. BOW, *Shape2*, SIFT (h) Loc. BOW, *Shape2*, GB (i) Loc. BOW, *Shape2*, *k*-AS

Figure 3.4: Localized bag-of-words classification accuracies for SIFT, GB, and *k*-AS, for *Caltech-256* (first row), *Shape* (second row), and *Shape2* (third row). SIFT-plots include the best GLOH and SC curves.

classification for *Caltech-256* (first row, plots (a) to (c)), *Shape* (second row, plots (d) to (f)), and *Shape2* (third row, plots (g) to (i)).

Caltech-256. Remarkably, for *Caltech-256* (see Figures 3.4 (a) to (c)), the clear ordering of detectors that is present for cluster precision and Naïve Bayes tends to dissolve. For different descriptors, highest scores are achieved for different detectors. Appearance-based SIFT and GLOH perform equally well with HarLap respectively HesLap, and equal to SalReg-SC. Most remarkably, shape-based GB outperforms SIFT and GLOH with HesLap and SalReg, in particular for high values of g. Adding location information boosts the performance of SalReg-GB by 20%. 3-AS gain 26%. Rotation invariant 3-AS perform equally well as HesLap-GLOH for $g = 9$. For 20 and 40 classes, k-AS loose performance relative to other features, which may be attributed to the weakness of the location model. The performance boost due to added location information decreases, but remains more important than the choice of descriptor, and at least as important as the choice of detector.

Shape. This tendency fully transfers to *Shape* (Figures 3.4 (d) to (f)). We observe again that shape-based features benefit to a large extent from location information: 45% boost for rotation invariant 2-AS, 36% for HarLap-GB, 35% for HesLap-SC, compared to 29% for appearance-based HesLap-SIFT. Further, the performance boost for shape-based features in response to increased location information even applies to the more challenging *Shape2* data set (Figures 3.4 (g) to (i)), where we perform the transition from *Shape*'s best-case scenario to more realistic images. HesLap-GB gains 21%, HesLap-SIFT 7%.

While for *Shape*, boosting over localized bag-of-words lifts the discriminant power of all feature types to a comparable level for $g = 9$ (Figures 3.4 (d) to (f)), shape-based features win for *Shape2*: HesLap-GB and rotation invariant 2-AS are best (42% respectively 44% accuracy). The best performing SIFT and GLOH combinations obtain 33% (HesLap-SIFT) and 31% (HesLap-GLOH). HarLap-SC obtains 35%.

Summary. To summarize, the localized bag-of-words results suggest two conclusions: first, adding location information can have a much bigger impact on classification accuracy than the choice of detector respectively descriptor. Second, shape features (Geometric Blur, Shape Context, k-Adjacent Segments) can benefit more from location than appearance-based ones. In particular, this boost can be sufficient for outperforming appearance-based features.

3.7 SUMMARY AND CONCLUSIONS

In this chapter, we have presented an evaluation of local shape- and appearance-based features. Building upon a formerly proposed method (Mikolajczyk *et al.*, 2005a) based on the comparison of clusterings, we measured local feature statistics over two complementary data sets representing shape- and appearance-based object

classes, respectively. We additionally contrasted two clustering techniques, and further evaluated local features as part of two general recognition frameworks with and without spatial information.

The key findings are: Local shape- and appearance-based features do not show great differences in terms of feature statistics over both shape- and appearance-based data sets. The choice of detector is more important on average than the choice of descriptor. Hessian-Laplace with SIFT and GLOH is best on average. Shape-based features (Geometric Blur, k-Adjacent Segments) perform mostly worse than appearance-based ones for classification based on simple occurrence statistics, but benefit more from added location information, and can even overtake appearance-based features on both shape- and appearance-based data.

FUNCTIONAL OBJECT CLASS DETECTION 4

Contents

Recognizing objects according to functional categories can be interpreted as a specific flavor of knowledge transfer, in which knowledge related to functional aspects is shared between objects belonging to different basic-level categories (Rosch *et al.*, 1976). In this chapter, we focus on a relatively narrow range of functional aspects, namely, different variants of grasping actions supported by objects, and demonstrate their transferability between instances of several basic-level categories. While our model is clearly limited with respect to the functional aspects that can be represented, it demonstrates the applicability of functional categorization to real world images of a standard benchmark data set. The particular choice of visual feature representation is motivated by the results of the experimental evaluation of Chapter 3, namely, the good performance of shape-based *k*-AS features (Ferrari *et al.*, 2008) in connection with a spatial model.

4.1 INTRODUCTION AND RELATED WORK

In recent years, computer vision has made tremendous progress in the field of object category detection. Diverse approaches based on local features, such as simple bag-of-words methods (Csurka *et al.*, 2004) have shown impressive results for the detection of a variety of different objects. More recently, adding spatial information has resulted in a boost in performance (Lazebnik *et al.*, 2006), and combining different cues has even further pushed the limits. One of the driving forces behind object category detection is a widely-adopted collection of publicly available data sets (Everingham *et al.*, 2010; Griffin *et al.*, 2007), which is considered an important instrument for measuring and comparing the detection performance of different methods. The basis for comparison is given by a set of rather abstract, basic level categories (Rosch *et al.*, 1976). These categories are grounded in cognitive psychology,

Figure 4.1: Basic level (left) vs. functional (right) object categories.

and category instances typically share characteristic visual properties.

In the context of embodied cognitive agents, however, different criteria for the formation of categories seem more appropriate. Ideally, an embodied, cognitive agent (an autonomous robot, e.g.), would be capable of categorizing and detecting objects according to potential uses, and w.r.t. their utility in performing a certain task. This *functional* definition of object categories is related to the notion of *affordances* pioneered by Gibson (1977).

Fig. 4.1 exemplifies the differentiation between functional and basic level categories, and highlights the following two key properties: 1) functional categories may generalize across and beyond basic level categories (both a mug and a watering-can are *handle-graspable*, and so is a hammer), and 2) basic level categories can be recovered as *composite* functional categories (a mug is both *handle-graspable*, *sidewall-graspable*, and can be *poured* from).

Attempts to detect objects according to functional categories date back to the early days of computer vision. Winston *et al.* (1983) were among the first to suggest functional characterizations of objects as consequences of basic geometric properties. Stark and Bowyer (1991) pioneered a body of work on functional categorization of CAD-inspired face-vertex object descriptions by geometric reasoning, and was later extended by visual input for recognizing primitive shapes from range images of idealistic scenes (Stark *et al.*, 1993). Rivlin *et al.* (1995) introduced an explicit mapping between geometric and corresponding functional primitives and relations, again restricted to a small class of parametric shapes. Bogoni and Bajcsy (1995) added force feedback for distinguishing among different tools that afford piercing other objects. Only recently, Saxena *et al.* (2007) stepped into the direction of more realistic settings, recognizing previously unseen, real world objects, but specifically tailored towards grasp point prediction. The approach is based on training a logistic regression model on annotated synthetic images, combining 2D filter responses with 3D range data in a dense, multi-scale image representation.

In this chapter, we approach the challenge of functional object categorization from a completely different angle. First, we build our system on robust and well-established grounds in the field of object recognition, applicable to real-world images of cluttered scenes. We explore the capabilities of a widely adopted detection framework, based on a suitable geometric local feature representation. Second, we choose to acquire functional category representations by observing few prototypical human-object interactions rather than explicitly modeling physical object properties. Naturally, the set of functional categories that our local feature-based vision system will be able to represent is restricted to those that are characterized by distinct visual

features. As an example, consider the bent shape of a mug handle, which suggests to grasp the mug in a specific way. We call such distinct visual features *affordance cues*, and base our system for functional object category detection on the recognition of these cues. In particular, this chapter makes the following contributions:

1. We present an integrated system for the acquisition, learning and detection of functional object categories based on affordance cues.
2. The system is based on a state-of-the-art object category detection framework, and acquires affordance cues from observing few prototypical human-object interactions.
3. We report first results for the detection of two functional object categories learned by our system, and demonstrate their generalization capabilities across and beyond basic level categories. We show that our system supports the interpretation of these categories as composite functional ones.

The rest of this chapter is organized as follows: Sec. 4.2 describes tutor-driven affordance cue acquisition. Sec. 4.3 presents the integration of affordance cues into a state-of-the-art object recognition framework. We give experimental results in Sec. 4.4, showing that promising detection performance can be achieved even for learning from as few as a single image, and conclude with a perspective on future work in Sec. 4.5.

4.2 AFFORDANCE CUE ACQUISITION

Given an observed human-object interaction featuring a single affordance cue (a video sequence plus tutor guidance), the purpose of the affordance cue acquisition sub-system is to obtain a visual feature-based representation of that cue. It proceeds by first estimating an accurate per-pixel segmentation of the *interaction region* (the region where tutor and object pixels overlap during interaction), and then extracting features in a local neighborhood around that region. Tutor guidance informs the system about the beginning and the end of an interaction. Fig. 4.2 gives an overview of affordance cue acquisition, which is detailed in the following.

4.2.1 Foreground/background segmentation and skin labeling

We employ the *Background Cut* (Sun *et al.*, 2006) algorithm originally proposed in the context of video conferencing for foreground/background segmentation. It combines global and local Gaussian mixture color models with a data-dependent discontinuity penalty in a Conditional Random Field model (Lafferty *et al.*, 2001), and provides accurate segmentations in near real-time.

In order to distinguish human tutor and manipulated object, we apply a likelihood ratio test on all pixels labeled as foreground by foreground/background segmentation. We build the ratio between the likelihood of a pixel originating from

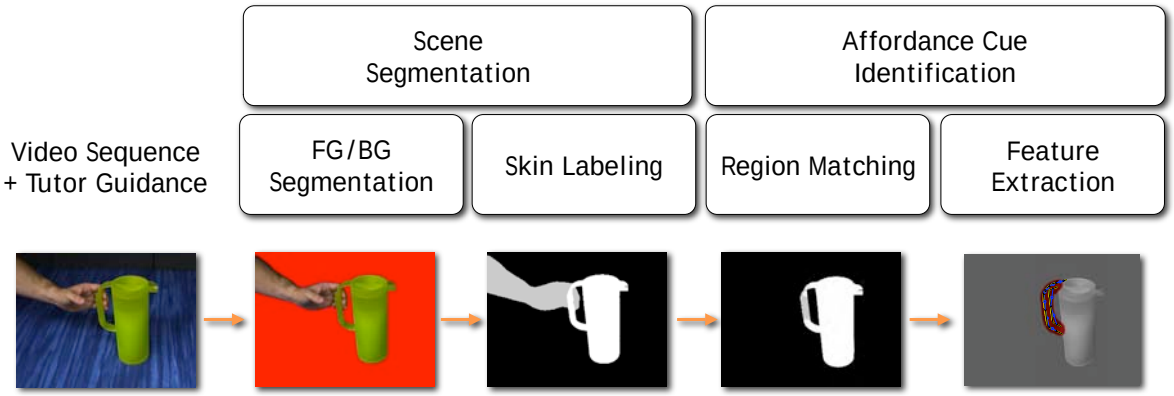


Figure 4.2: Affordance cue acquisition overview.

object color and the corresponding likelihood for skin color, using a pre-trained skin color model (Jones and Rehg, 1999). Fig. 4.2 includes an example labeling (black denotes *background*, white *object*, and gray *skin*).

4.2.2 Region matching

We determine the interaction region as the set of object pixels that has been occluded by the human tutor in the course of an interaction. We identify those pixels by choosing two frames from the interaction sequence, i) one during the interaction, and ii) one after (but with the object still visible). Then, the set of occluded object pixels is computed as the intersection of all skin-labeled pixels of frame i) with all object-labeled pixels of frame ii), transformed in order to equalize object pose differences between the two frames. The transformation is obtained by estimating the homography between frames i) and ii), using RANSAC. Initial point-to-point correspondences are established by Robust Nearest-Neighbor Matching of SIFT descriptors (Lowe, 2004) on Harris-Laplace interest points (Mikolajczyk *et al.*, 2005b) (see Fig. 4.3 and Fig. 4.4 for additional examples).

4.2.3 Feature extraction

Our representation of affordance cues is based on geometric local features called k-Adjacent Segments (k-AS) (Ferrari *et al.*, 2006a), initially proposed in the context of shape-matching line drawings to real images (Ferrari *et al.*, 2006b). k-AS detect distinct edge segments in an image, form groups of k such segments, and then encode their relative geometric layout in a low dimensional, scale-invariant shape descriptor. In our experiments, we consistently use $k = 2$, since 2-AS features have shown a good discrimination/repeatabilty tradeoff in our experimental evaluation in Chapter 3. We augment the groups returned by the 2-AS detector by additional

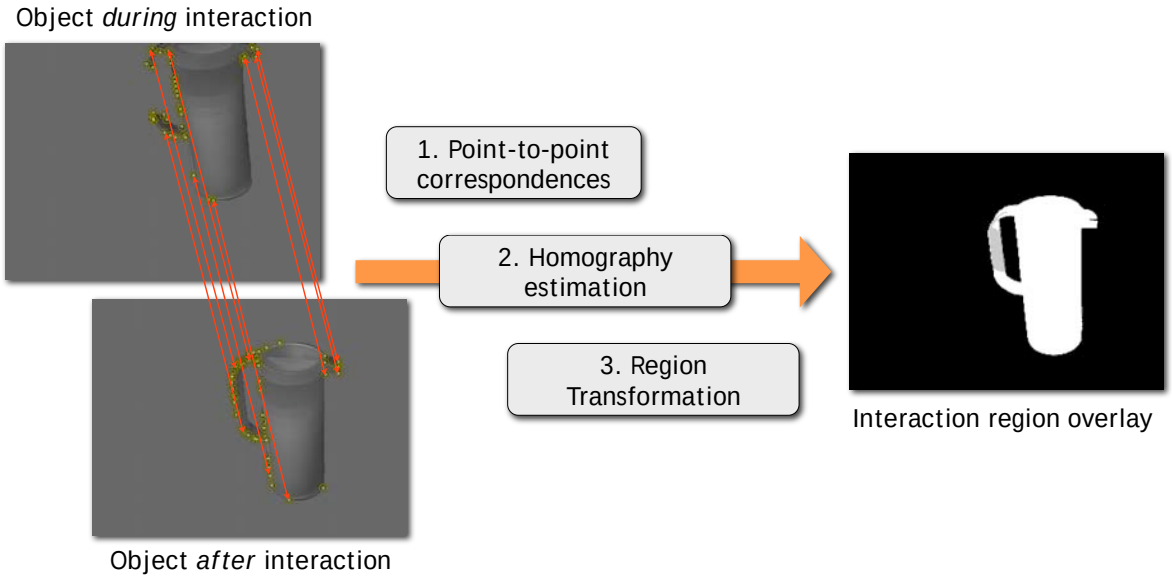


Figure 4.3: Region matching overview.

Figure 4.4: Region matching examples.

pairs of edge segments according to perceptual grouping criteria in the spirit of Zillich (2007). By indexing the space of detected edge segments by histogramming orientations, we maintain linear grouping complexity. We transfer the per-pixel segmentation into 2-AS features by growing the interaction region, and including edge segments with sufficient overlap.

4.3 AFFORDANCE CUE-BASED OBJECT DETECTION

A variant of the Implicit Shape Model (ISM) (Leibe *et al.*, 2006a) serves as the basis for our functional object category detection system, using the affordance cue representation of Sec. 4.2. We extend the original model in order to allow for independent training of several different affordance cues, and flexible combination for detecting composite functional categories. Fig. 4.5 gives an overview of the ISM.

Training. Training an ISM for an affordance cue amounts to matching acquired affordance cue features to a previously built codebook, and storing the relative position (x, y) and size (*scale*) of the object w.r.t. the feature occurrence along with matched codebook entries. Position and scale can be easily obtained from a bounding box surrounding all object-labeled pixels from the acquisition stage.

Detection. For detecting an affordance cue in a previously unseen image, all features in a test image are again matched to the codebook. For every matched codebook entry, each stored feature occurrence probabilistically votes for a hypothesized object position in a generalized three-dimensional Hough voting space $(x, y, scale)$. The probabilistic vote is a function of its distance to the codebook entry in feature space, the edge-strength of the corresponding image feature, and the amount of overlap between the stored feature occurrence and the interaction region of the originating training affordance cue.

We estimate modes in the distribution of votes by standard kernel density estimation techniques, and accept modes as detections according to a confidence threshold. Since we are interested in a precise estimate of where exactly an affordance cue is located in an image, we proceed by back-projecting those features into the image, which contribute significantly to either of the modes, by selecting a fixed volume of probability mass around each mode. Fig. 4.8 shows example detections, where highlighted edges correspond to back-projected 2-AS features.

Combining Multiple Affordance Cues. One of the reasons for choosing an ISM for our approach is its extendibility to multiple affordance cues. Having trained multiple affordance cue models separately, these models can be joined for detecting *composite functional categories* by combining votes of all models in a single Hough voting space, and estimating modes of the joint distribution.

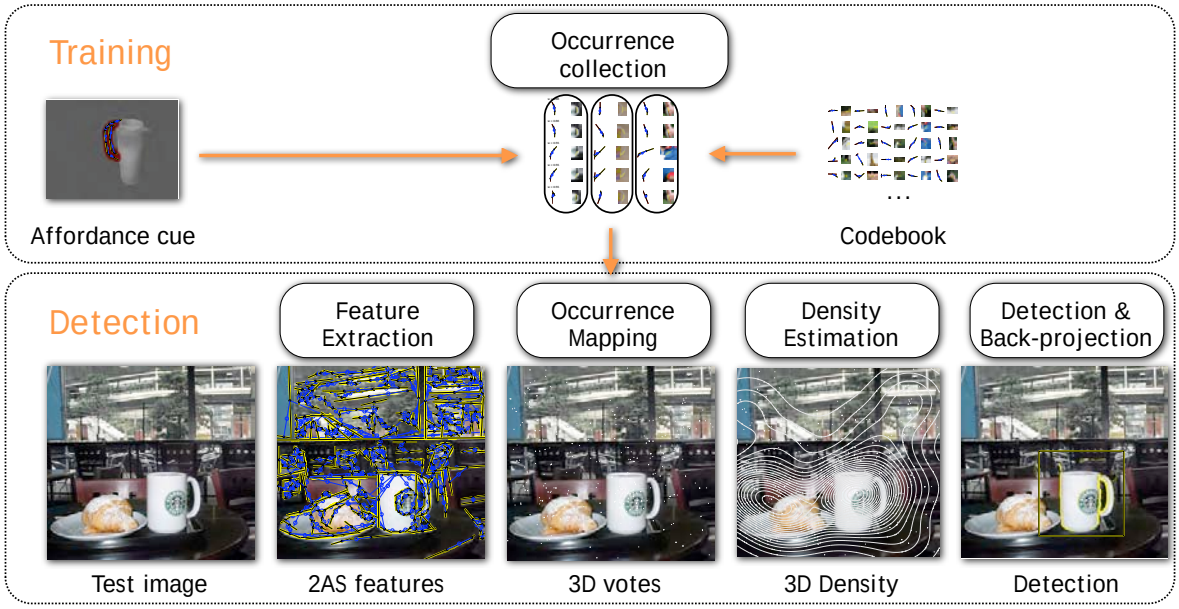


Figure 4.5: Implicit Shape Model Overview.

4.4 EXPERIMENTAL RESULTS

For all experiments, we build generic codebooks by hierarchical agglomerative clustering of 2-AS features from a set of randomly selected real images, augmented by additional pairs of edge segments according to perceptual grouping criteria, as presented in Sec. 4.2. We report qualitative results for the detection performance of our system on a subset of the ETHZ Shape Classes data set (Ferrari *et al.*, 2006b), and a series of images from the environment of our embodied, cognitive agent. Example detections are given in Figure 4.8. Each row corresponds to a single experiment, unless otherwise stated. For each row (a) to (g), columns (2) to (5) give example detections for a system that has been trained solely on the highlighted affordance cue features in column (1). Line segments are plotted in yellow, and pairs selected as 2-AS feature are connected by a blue line. Row (d) continues example detections of row (c), and row (g) depicts detections from a system trained on affordance cue features (1)(a) *and* (1)(f). Back-projected edges from the *handle-graspable* detector are plotted in yellow, those from the *sidewall-graspable* detector in red.

The *handle-graspable* category. We begin by giving results for the *handle-graspable* functional category (rows (a) to (c) of Fig. 4.8), learned from affordance cue features of single images given in column (1). We observe that the models learned from either of the three mugs perform comparably well in detecting handle-like structures in the given test images, despite apparent appearance differences between the objects used for training and testing, and considerable background clutter. The affordance cue features learned from mugs (1)(a) and (1)(c) achieve slightly more accurate localization of handle-like features in the test images, apparently due to

Figure 4.6: Comparison of affordance cue-based (*handle-graspable*) vs. whole-object training. (1) and (3) depict detections, (2) and (4) the corresponding top-five 2-AS features, stored codebook occurrences, and matched codebook entries (from left to right).

their symmetry w.r.t. the horizontal axis, resulting in increased repeatability.

Row (d) highlights the generalization of a *handle-graspable* model learned from mug (1)(c) over other object categories such as *coffee pot*, *vase*, and *electric water jug*. Image (5)(d) indicates the limitations of our approach. While the detector mistakenly fires on a circular sign in the background (false positive), it misses an obvious *handle-graspable* affordance cue on the white Thermos bottle (false negative). While the false positive can be explained by the limited information encoded by the 2-AS features, the false negative may be attributed to predominant background structures.

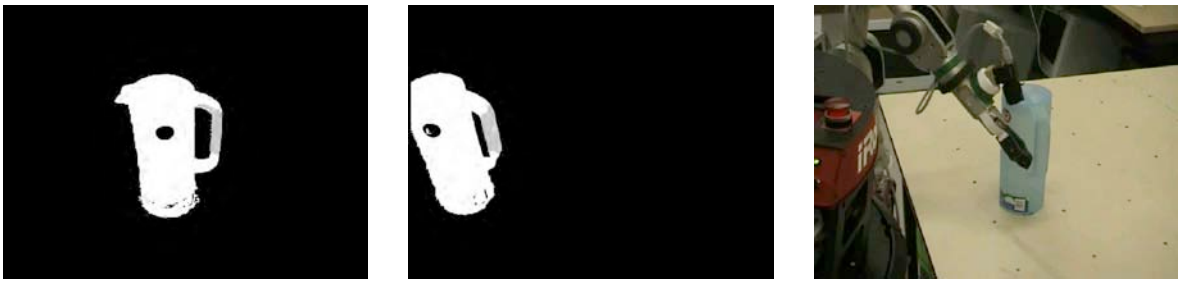
Affordance cues for feature selection. An interesting question is how the performance of object detection based on affordance cues compares to the performance of a system that has been trained without being directed towards these cues. Fig. 4.6 contrasts detections of an ISM trained on *handle-graspable* affordance cue features (1) vs. an ISM trained on *all* features of a mug (3). In fact, the latter provides less accurate localization of the mug in the test image; none of the shown three bounding boxes in (3) (the three most significant modes) comes as close to the true position of the mug as the single one in (1). Fig. 4.6 (2) and (4) pinpoint the difference, by listing the respective top-five features contributing to the most significant mode, together with the corresponding stored codebook occurrences and matched codebook entries. While the *handle-graspable* detector correctly relates handle features from training and test image, the *all*-detector matches mostly texture features between the two, misleading its prediction of object center and scale.

One possible approach to overcoming the weak discriminative power of the employed features is the exploitation of additional affordance cues in a joint, composite functional category model, as will be demonstrated in the following.

The *sidewall-graspable* category. Rows (e) and (f) of Fig. 4.8 show the detection results for a second category, *sidewall-graspable*, again learned from single images. In row (e), a model has been learned from a bottle, and from a mug in row (f). The *sidewall-graspable* detector exhibits remarkable performance in the detection of sidewall-like structures in cluttered images, although it is slightly more susceptible

to false positives than the *handle-graspable* detector, again due to the limitations of the employed features (see (e)(5)).

The *handle-graspable/sidewall-graspable* category. We now combine both *handle-graspable* and *sidewall-graspable* affordance cues by training two independent ISM models, one for each cue, and joining their predictions for detection. In fact, the combination of both cues improves the detection performance of our system (example detections are given in row (g)). In particular, the *sidewall-graspable* affordance cue compensates for inaccuracies in the localization of *handle-graspable* features. By back-projecting features, the joint detector is able to to distinguish and accurately localize both of the two affordance cues, shown in yellow (*handle-graspable*) and red (*sidewall-graspable*).



Left camera Right camera Grasping attempt

Figure 4.7: Binocular affordance cue acquisition and resulting grasping attempt.

Grasping. Fig. 4.7 depicts an attempt to grasp a jug at its handle by a robot arm mounted onto our agent, after the system has acquired the corresponding affordance cue. Although actual grasping fails due to limited visual servoing capabilities, the agent manages to touch the jug at the correct position. We applied affordance cue acquisition independently for two cameras of a calibrated stereo rig, and obtained 3D coordinates by triangulation.

4.5 CONCLUSIONS AND FUTURE WORK

In this chapter, we have approached the challenge of functional object categories from the perspective of state-of-the-art object detection, and presented a system for the tutor-driven acquisition, learning, and recognition of affordance cues in real-world, cluttered images. Clearly, our system is limited by the 2D nature of the used local features, but exhibits promising detection performance in our experiments even for one-shot learning.

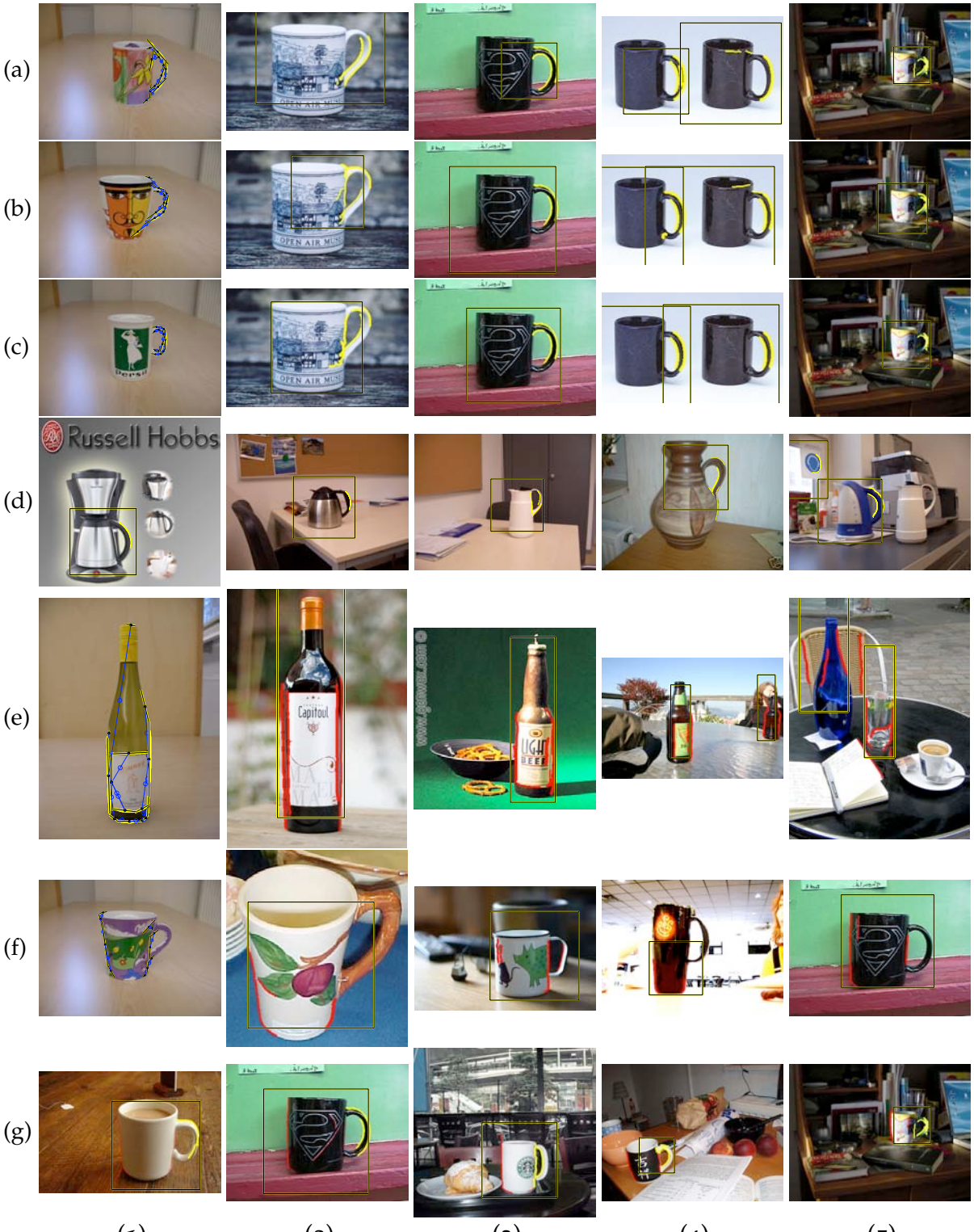


Figure 4.8: Example detections (see Section 4.4 for details).

5 A SHAPE-BASED OBJECT CLASS MODEL FOR KNOWLEDGE TRANSFER

Contents

RETURNING to the introductory example of Section 1.1, this chapter pursues the question how to reuse an available object class model in order to facilitate the learning of a new one. For this purpose, it designs a shape-based object class model for knowledge transfer, and demonstrates, e.g., the successful reuse of an available *horse* model for the learning of *giraffe* and *swan* models. The suggested model is compositional, in that it allows to transfer knowledge restricted to a subset of components, and incremental, meaning that an existing model can be further specialized by adding more training examples. In contrast to the object class model presented in Chapter 4, transferable knowledge is not restricted to groups of local shape features, but can additionally encompass their spatial layout and pair-wise symmetry relations. While the particular choice of local shape feature is again inspired by the results of the experimental evaluation of Chapter 3, this chapter proposes a novel flavor of local shape feature as an extension.

5.1 INTRODUCTION

Object class detection has made impressive progress in recent years. Most models rely on robust local features and powerful learning approaches such as SVMs

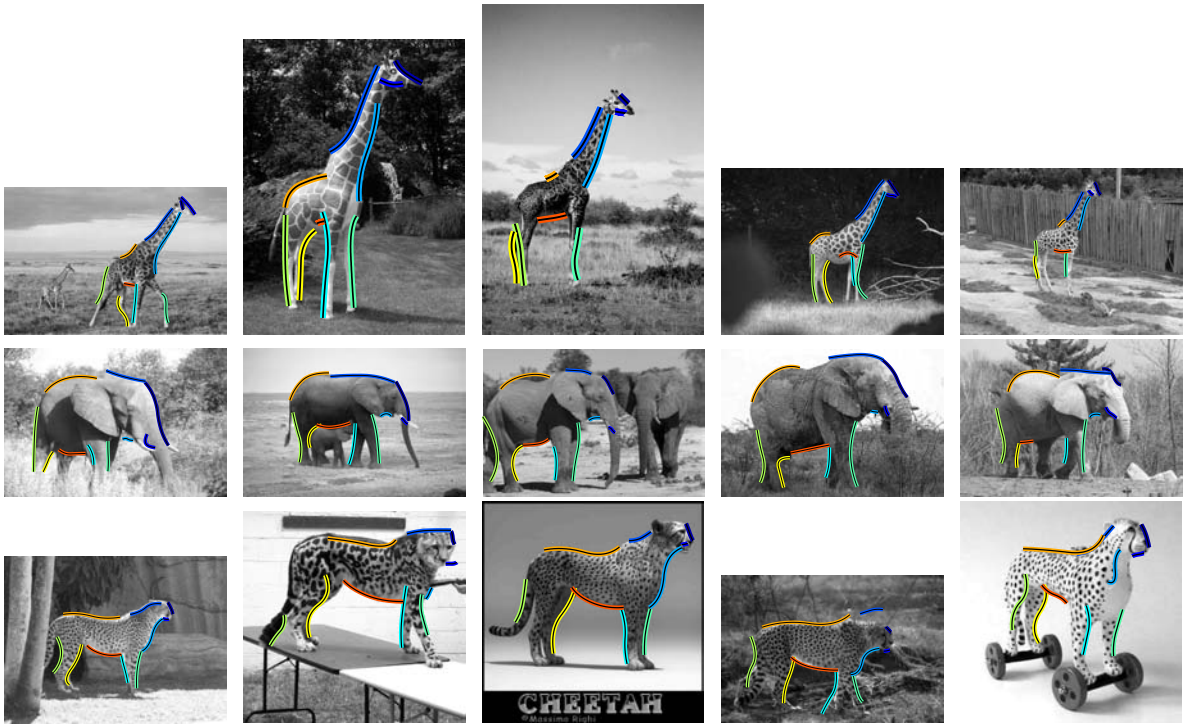


Figure 5.1: Animal detections using *1*-shot models.

requiring substantial amounts of training data per object class. In order to scale to larger numbers of object classes than it is possible today it is widely believed that information from one or a set of object classes should be transferred to the learning of new object classes. This would reduce the required training data for new object classes and might even facilitate *1*-shot learning of novel classes. This idea of transfer learning has long been argued for both from a psychological point of view (Ahn and Brewer, 1993; Moses *et al.*, 1993) as well as from a computer vision point of view (Bart and Ullman, 2005b; Fei-Fei *et al.*, 2006; Ferencz *et al.*, 2005; Fink, 2004; Levi *et al.*, 2004; Miller *et al.*, 2000; Zweig and Weinshall, 2007). While these approaches have shown to enable object class learning from small numbers of training instances none of these models—as of today—has reached wide-spread use.

The starting point and main contribution of this chapter is therefore to take a fresh look at the problem and to explicitly design a novel object model that directly lends itself to transfer learning. We start with the observation that there are at least three different types of knowledge that should be transferable between object models. First, the appearance or shape of an object part should be transferable (e.g., the shape of a leg or wheel). Second, local symmetries between parts are often shared by different object classes (e.g., the symmetry between front- and back-legs for quadrupeds). And third, the layout of the different parts is often at least partially shared among different object classes (e.g., the layout of head, torso and body for quadrupeds and birds, see also Sec. 5.5.2). In the following, we devise a part based model with a separate factor for each of these properties that allows, e.g., to

transfer the layout of an object model either fully or only partially, constrained to an appropriate subset of object parts. The main contributions of this chapter are:

- We propose a novel *shape-based object model for knowledge transfer* that can be factored into per-part components and enables transfer of full or partial knowledge.
- We demonstrate the *importance of symmetries*, a primitive rarely used for object detection, for both, object model learning as well as knowledge transfer.
- We experimentally validate our object model on the ETHZ Shape Classes data set, demonstrating competitive performance with prior work.
- We demonstrate that our model enables transfer of information on a quadrupeds database where we transfer the full layout and symmetry information. In addition, we also show successful partial information transfer in two interesting and quite different cases.

The remainder of this chapter is organized as follows: After a review of related work we first introduce our model (Sect. 5.2) and validate its performance (Sect. 5.3). We then describe the knowledge transfer approach (Sect. 5.4) and demonstrate results for full and partial model transfer (Sect. 5.5). We conclude with an outlook on future work (Sect. 5.6).

5.1.1 Related work

Transferring knowledge such as appearance, shape or symmetries between object classes is an important topic due to its potential to enable efficient learning of object models from a small number of training examples. It provides the basis for scalability to large numbers of classes. Broadly speaking, related work in knowledge transfer falls into three different categories: *distance metric learning*, *joint learning of multiple object classes*, and *use of prior information*.

The main idea of *distance metric learning* is to learn a representation for a set of a priori known classes in the form of a distance metric among them (Fink, 2004; Thrun, 1996). This metric can then be used directly to classify instances of an unknown class. Bart and Ullman (2005a) replace features from known classes with ones from a new class, implicitly re-using the learned distance metric. These approaches have shown to improve *1*-shot learning mainly for simple objects and handwritten chars.

In the context of *joint learning of multiple object classes*, machine learning has developed the notion of multiple task learning. This allows learners to benefit from the similarity of multiple, different learning tasks (Ben-David and Schuller, 2003). A second line of research is based on joint training of multiple classifiers, which draw from a common pool of features (Amit *et al.*, 2007; Torralba *et al.*, 2004), thereby enabling feature sharing. While these approaches clearly reduce the amount of necessary training data per object class, knowledge transfer happens rather implicitly. Explicit and controlled transfer of knowledge is not supported.

The *use of prior information* is most related to this work, and comes in multiple flavors. Levi *et al.* (2004) use models of unrelated object classes to prime feature selection of an unknown class. Bart and Ullman (2005b) directly use similarities

to known classes to represent new classes. Zweig and Weinshall (2007) propagate knowledge along a hierarchy, by learning and combining classifiers for individual levels of the hierarchy to yield a more effective classifier for specific leaf classes. Fei-Fei *et al.* (2006) transfer information via a Bayesian prior on object class models, using knowledge from known classes as a generic regularizer for newly learned models.

While this chapter clearly falls into the last category, we stress the following key differences from related work: most importantly, our approach is designed to allow an explicit, controlled transfer of prior knowledge. In particular, it facilitates knowledge transfer at both the level of a full model, and selected aspects of a model. Being based on an assembly of parts, their spatial layout and symmetry relations provides a rich source of independently transferable properties, ranging from fairly general (overall spatial layout of parts) to very specific (local part shape). We strongly believe that both the explicit as well as the partial transferability of prior information are key ingredients to make knowledge transfer a common tool for object class modeling and learning.

Concerning object recognition, our work is most related to part-based methods such as the Constellation Model (Fergus *et al.*, 2003) or the Implicit Shape Model (Leibe *et al.*, 2006a). While the non-parametric scene-object-part model of Sudderth *et al.* (2008) requires less supervision than ours, its appearance-based, visual word part representation is limited compared to our flexible combination of local shape and semi-local symmetry relations. Zhu (1999) give a fundamental treatment of probabilistic shape modeling and Gestalt principles, including symmetry. Park *et al.* (2008) evaluate the accuracy of several symmetry detection algorithms. In contrast to early attempts (Brooks, 1983; Nevatia and Binford, 1977), this chapter shows the successful application of a particular kind of symmetry relations (Brady and Asada, 1984; Saint-Marc *et al.*, 1993) to object class detection in real images.

5.2 THE MODEL

Our model is inspired by the *Constellation Model* (Fergus *et al.*, 2003), but goes beyond this model in several ways. First, it relies entirely on shape information. Second, we propose a Data-Driven Markov Chain Monte Carlo (DDMCMC) (Zhu *et al.*, 2000) technique for efficient inference, which increases the number of features the system can handle by several orders of magnitude. Third, we enrich the original formulation comprising object parts, their relative scales, and spatial layout, by pair-wise symmetry relations between parts. Pair-wise relations even between simplistic line features have proven to be powerful cues for recognition (Leordeanu *et al.*, 2007), which we confirm in our experiments. Fourth, we demonstrate that knowledge can be effectively transferred between different model instances, on two different levels.

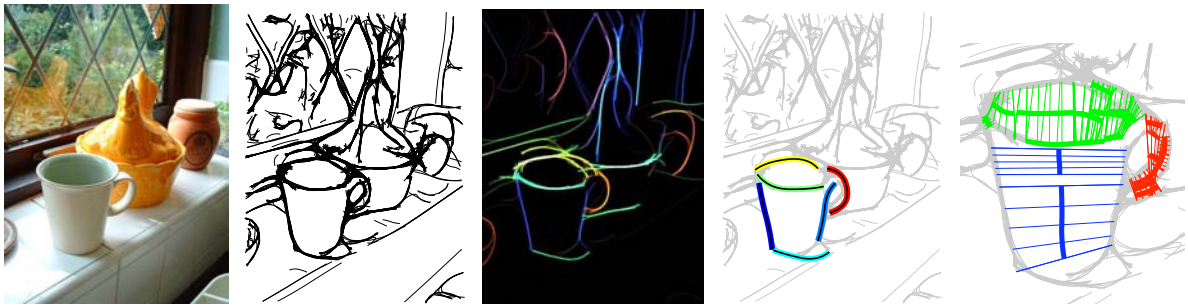


Figure 5.2: From left to right: Original image, local shape features, color-coded part likelihoods, detection hypothesis, selected symmetry lines and axes.

5.2.1 Local shape features

We introduce a novel flavor of local shape features, which constitute a discrete, over-complete representation of image contours. The shape features are based on the *Contour Segment Network* (CSN) (Ferrari *et al.*, 2006b), and its associated local companions, k-Adjacent Segments (k-AS) (Ferrari *et al.*, 2007). We suggest important additions to these techniques, as detailed below.

Starting from an edge probability map of the Berkeley natural boundary detector (Martin *et al.*, 2004), a collection of discrete, roughly straight contour segments is formed, and subsumed in a network topology (the CSN), based on spatial proximity and edge continuity constraints. Since, by design, the CSN can be assumed to provide an over-segmentation of image edges, meaning that object parts are likely to be fragmented into several segments, we *simultaneously* include k-AS with $k \in \{1, \ldots, K\}$ into our representation, to increase the chance of having a shape feature available that matches one-to-one to an object part. In practice, we use $K = 5$.

Further, we *unify* the representation of k-AS for varying k by fitting a parametric B-spline curve to all constituent edgel chains, using the exact same parameterization, independent of k. This offers the additional benefit of retaining the original *curvature* information and increasing the discriminative power of the features compared to the original k-AS represented by straight line approximations.

In our implementation, we first transform all constituent edgel chains of a given k-AS into a translation and scale invariant space, using Procrustes analysis (Cootes, 2000). We use the resulting spline parameters as a low-dimensional local shape description. In all experiments, we use pairs of quadratic B-splines, resulting in an 8-dimensional descriptor. We prune the set of features based on the goodness of fit of the splines. Fig. 5.2 shows all 1640 local shape features of an image.

5.2.2 Semi-local symmetry relations

As shown in the literature (Ferrari *et al.*, 2006b), and confirmed by our experiments in Chapter 3, local shape features based on contour segments tend to be more generic in nature than local texture features, and hence provide relatively weak discrimination

among object parts and background clutter, if used in isolation. We therefore include another powerful perceptual cue into our model, which relates pairs of local shape features by identifying and describing *symmetries* between them. In particular, we use a B-spline-based implementation (Saint-Marc *et al.*, 1993) of *Smoothed Local Symmetry* (SLS). SLS were originally proposed by Brady and Asada (1984) in the context of planar shape analysis.

SLS relate two parametric shapes by determining pairs of points that fulfill a *local symmetry* constraint: A point p_1 on shape s_1 is locally symmetric to a point p_2 on s_2, if the respective angels between the connecting line between p_1 and p_2, and the normal vectors at p_1 and p_2, are equal. The set of all locally symmetric point pairs and their associated connecting lines (the *symmetry lines*) then define the *symmetry axis* between the shapes: it consists of the mid-points of the symmetry lines. Fig. 5.2 (right) depicts several selected symmetry lines and axes between local shape features of a mug (blue: symmetries between side-wall features, green: between rim features, red: between handle features).

Starting from the spline-based representation of SLS, we now devise a semi-local symmetry descriptor, which captures both the shape of the symmetry axis and the lengths of the symmetry lines, in order to characterize the symmetry. The first is achieved by representing the axis as a local shape feature, exactly as described in Sect. 5.2.1. We compute a fixed number of symmetry lines (usually 10) and record a profile of their respective lengths, as we traverse the symmetry axis from end to end. We then reduce the dimensionality of the resulting length profile vector by PCA (usually to 3). Fig. 5.7 (b) depicts length profiles as bar plots corresponding to the symmetry axes denoted by gray lines in Fig. 5.7 (a).

5.2.3 Probabilistic model

We now describe the probabilistic model that subsumes individual part shapes S, binary symmetry relations B, relative part scales R, and their overall spatial layout X. We borrow from the notation of Fergus *et al.* (2003) where appropriate.

During detection, our goal is to find an *assignment* of all P model parts to local shape features, which we denote the detection hypothesis $H = (h_1, \ldots, h_P)$. That is, h_p contains a local shape feature identifier assigned to part p. We formulate the detection problem as a maximum a posteriori hypothesis search over the distribution $p(X, R, S, B, H|\theta)$, which is the joint posterior distribution of H and image evidence, given a learned model θ. It factors as follows:

$$p(X, R, S, B, H|\theta) = \underbrace{p(S|H,\theta)}_{\textit{Local Shape}} \underbrace{p(B|H,\theta)}_{\textit{Symm. Rel.}} \underbrace{p(X|H,\theta)}_{\textit{Layout}} \underbrace{p(R|H,\theta)}_{\textit{Rel. Scale}} \underbrace{p(H|\theta)}_{\textit{Prior}} \tag{5.1}$$

In all experiments, we assume a uniform prior $p(H|\theta)$.

Local Part Shape. Local part shape $S(h_p)$ is modeled by a Gaussian density on spline parameters (see Sect. 5.2.1).

$$p(S|H,\theta) = \prod_{p=1}^{P} \mathcal{N}(S(h_p)|\theta). \tag{5.2}$$

Binary Symmetry Relations. We instantiate the binary relation component of our model with a joint density over SLS descriptors, as described in Sect. 5.2.2. It comprises all pairs of parts, excluding self- and duplicate pairings. For each pair, it factors into two Gaussian densities, where one governs the SLS axis spline parameters $B_a(h_i,h_j)$, and one the PCA-projection of the corresponding symmetry line length profile $B_l(h_i,h_j)$.

$$\begin{aligned} p(B|H,\theta) &= \prod_{i=1}^{P-1} \prod_{j=i+1}^{P} p(B(h_i,h_j)|\theta) \\ p(B(h_i,h_j)|\theta) &= \mathcal{N}(B_a(h_i,h_j)|\theta)\, \mathcal{N}(B_l(h_i,h_j)|\theta) \end{aligned} \tag{5.3}$$

Spatial Layout and Relative Scales. We model the spatial layout of constituent model parts as a joint Gaussian distribution over their coordinates $X(H)$ in a translation- and scale-invariant space (the *constellation*), again using Procrustes analysis (Cootes, 2000). The model allocates independent Gaussians for the relative scale $R(h_p)$ of each part, i.e., the ratio between part and constellation scale.

$$p(X|H,\theta)\, p(R|H,\theta) = \mathcal{N}(X(H)|\theta) \prod_{p=1}^{P} \mathcal{N}(R(h_p)|\theta) \tag{5.4}$$

5.2.4 Learning and inference

Learning. We learn maximum likelihood model parameters θ for all model components using supervised training. Supervision is provided by labeling contour segments in training images (see Sect. 5.2.1), which in practice amounts to a few mouse clicks per object instance.

Inference. During detection, we search for $H_{\text{MAP}} = \arg\max_H p(H|X,R,S,B,\theta)$, the maximum a posteriori hypothesis. This is equivalent to $\arg\max_H p(X,R,S,B,H|\theta)$. We approximate H_{MAP} by drawing samples from $p(X,R,S,B,H|\theta)$ using the Metropolis-Hastings (MH) algorithm (Gilks *et al.*, 1996). We use the Single Component update variant of MH, since it allows to separately update individual components of the target density, conditioned on the remaining portion of the current state of the Markov chain. This opens the possibility to guide the sampling towards high density regions by data-driven, bottom-up proposals (Tu *et al.*, 2005; Zhu *et al.*, 2000). Similar to (Lee and Cohen, 2004), we define P independent proposal distributions of the form $q_p(S(h_p)|\theta) = \mathcal{N}(S(h_p)|\theta)$, based on the likelihoods of the local shape part

model. Fig. 5.2 depicts a joint, color-coded visualization of all part proposals for a mug model consisting of 7 parts (two side-walls, two rim parts, one bottom part, two handle parts), together with an example detection based on exactly these proposals. Notably, the combined part likelihood is much sparser than the corresponding visualization of all local shape features to the left of Fig. 5.2.

We obtain the following acceptance ratio for changing the current hypothesis $H = (H_{-p}, h_p)$ to $H' = (H_{-p}, h'_p)$, where H deviates from H' only in component h_p, and H_{-p} denotes the other components that are kept.

$$\alpha = \min \left\{ 1, \frac{p(X, R, S, B, h'_p | H_{-p}, \theta)\ q_p(S(h_p)|\theta)}{p(X, R, S, B, h_p | H_{-p}, \theta)\ q_p(S(h'_p)|\theta)} \right\} \tag{5.5}$$

Note that most of the terms in this ratio actually cancel due to the factorization of our model (namely the ones not involving the part under consideration p). This implies in particular that the number of pair-wise relations that have to be computed per iteration grows only linearly, and not quadratically, with increasing number of parts P. Further, since the sampling process is guided by data-driven proposals, the number of pair-wise relations considered is orders of magnitudes smaller than the number of all possible pairings. We exploit this fact by computing SLS in a lazy fashion, and subsequently caching them, which greatly improves runtime behavior. For a typical image with several thousands of features, our model typically (re-) considers at most a few tens of thousands of pairs, not tens of millions.

Detection. We detect object instances by running m independent Markov chains, and memorizing the per-chain highest-scoring hypotheses. In all experiments, we run $m = 50$ chains, for a maximum number of 1000 iterations, yielding runtimes of under a second per Markov chain. We use the greedy non-maximum suppression described in Fritz and Schiele (2008) to prune overlapping hypotheses.

5.3 SHAPE CLASSES EXPERIMENTS

We evaluate the performance of our model on a standard shape benchmark (Ferrari *et al.*, 2006b) and report detection results on 4 of the 5 classes of the ETHZ Shape Classes data set (see Fig. 5.1 and 5.3). We use the test protocol of Ferrari *et al.* (2007): experiments are conducted in 5-fold cross-validation. For each class, we learn 5 different models by sampling 5 subsets of half of the class images at random. The test set for a model then consists of all other images in the data set (taken from all 5 classes). Performance is measured as the average detection rate at 0.4 false positives per image (FPPI), measured for an overlap of 0.2 between ground truth and detection bounding boxes.

We compare the results with two methods that follow the exact same evaluation protocol, and which have been published prior to the work presented in this chapter. One is shape-based (Ferrari *et al.*, 2007), and one is based on topic-decompositions of HOG-like features (Fritz and Schiele, 2008). For (Ferrari *et al.*, 2007), we consider

Results	Bottle	Giraffe	Mug	Swan	*average*
Ferrari *et al.* (2007)	83.2 (7.5)	58.6 (14.6)	83.6 (8.6)	75.4 (13.4)	75.2 (11.0)
Fritz and Schiele (2008)	76.8 (6.1)	90.5 (5.4)	82.7 (5.1)	84.0 (8.4)	83.5 (6.3)
Our model	91.0 (3.8)	**91.7 (4.1)**	76.6 (9.9)	77.7 (5.8)	84.3 (5.9)
Our model, SLS	**94.4 (3.8)**	**91.7 (2.6)**	**84.5 (4.7)**	**88.8 (6.9)**	**89.9 (4.5)**
Ommer and Malik (2009)	89.3	77.3	91.8	95.7	88.5
Maji and Malik (2009)	96.4	**89.6**	**96.7**	88.2	92.7
Srinivasan *et al.* (2010)	**100.0**	**89.6**	93.6	**100.0**	**95.8**

Table 5.1: ETHZ Shape Classes results: average detection rates, standard deviations given in brackets where applicable. Comparison to prior (upper half) and more recent work (lower half); for each half, bold font marks the per-column best value.

the results based on *learned* models rather than the ones based on hand-drawings, as they are comparable to our approach. For the same reason we do not compare against Ravishankar *et al.* (2008); Zhu *et al.* (2008). We further relate our results to results that were reported more recently (Ommer and Malik, 2009; Maji and Malik, 2009; Srinivasan *et al.*, 2010), but obtained by following a slightly different and thus not fully comparable protocol (a single test run instead of 5-fold cross-validation, an overlap of 0.5 instead of 0.2).

As shown in the upper half of Tab. 5.1, our model without symmetry compares favorably to the previous results of Ferrari *et al.* (2007) and Fritz and Schiele (2008) on *bottles* and is slightly better on *giraffes*. For *mugs* however the performance is lower and for *swans* it is between Ferrari *et al.* (2007) and Fritz and Schiele (2008). On average it outperforms both methods. We attribute this good performance to the combination of robust local shape features with a flexible spatial model. Adding symmetry relations (SLS) significantly increases performance for two classes (11% for *swans*, 8% for *mugs*) and also slightly for *bottles* (3%). As a consequence our model performs better than both previous methods on all four classes. Using symmetries attains 89.9% on average, 6.4% better than the next best previous method.

As concerns the relation to more recent work (see the lower half of Tab. 5.1), our approach using symmetries performs on average on a similar level as the implicit shape model variant proposed by Ommer and Malik (2009) (modulo the differences in the evaluation protocol). The discriminatively trained ISM of Maji and Malik (2009) significantly improves over Ommer and Malik (2009) except for *swans*, and achieves higher average performance (on a single test run) than either of our models despite the tighter overlap criterion. The discriminatively trained contour model of Srinivasan *et al.* (2010) provides a further significant improvement of average performance, achieving perfect detection rates on *bottles* and *swans*, again outperforming our models.

As a general observation, we note that all three recent methods (Ommer and

Malik, 2009; Maji and Malik, 2009; Srinivasan *et al.*, 2010) rely on discriminative rather than representative techniques, either in the form of a verification step (Ommer and Malik, 2009) or as an integral part of the model design (Maji and Malik, 2009; Srinivasan *et al.*, 2010). We suspect that discriminative training is one of the key factors for the superior performance of these approaches, while the object class model presented in this chapter is strictly representative. Following this intuition, we introduce discriminatively trained part detectors as a replacement for representative local shape detectors in Chapter 7.

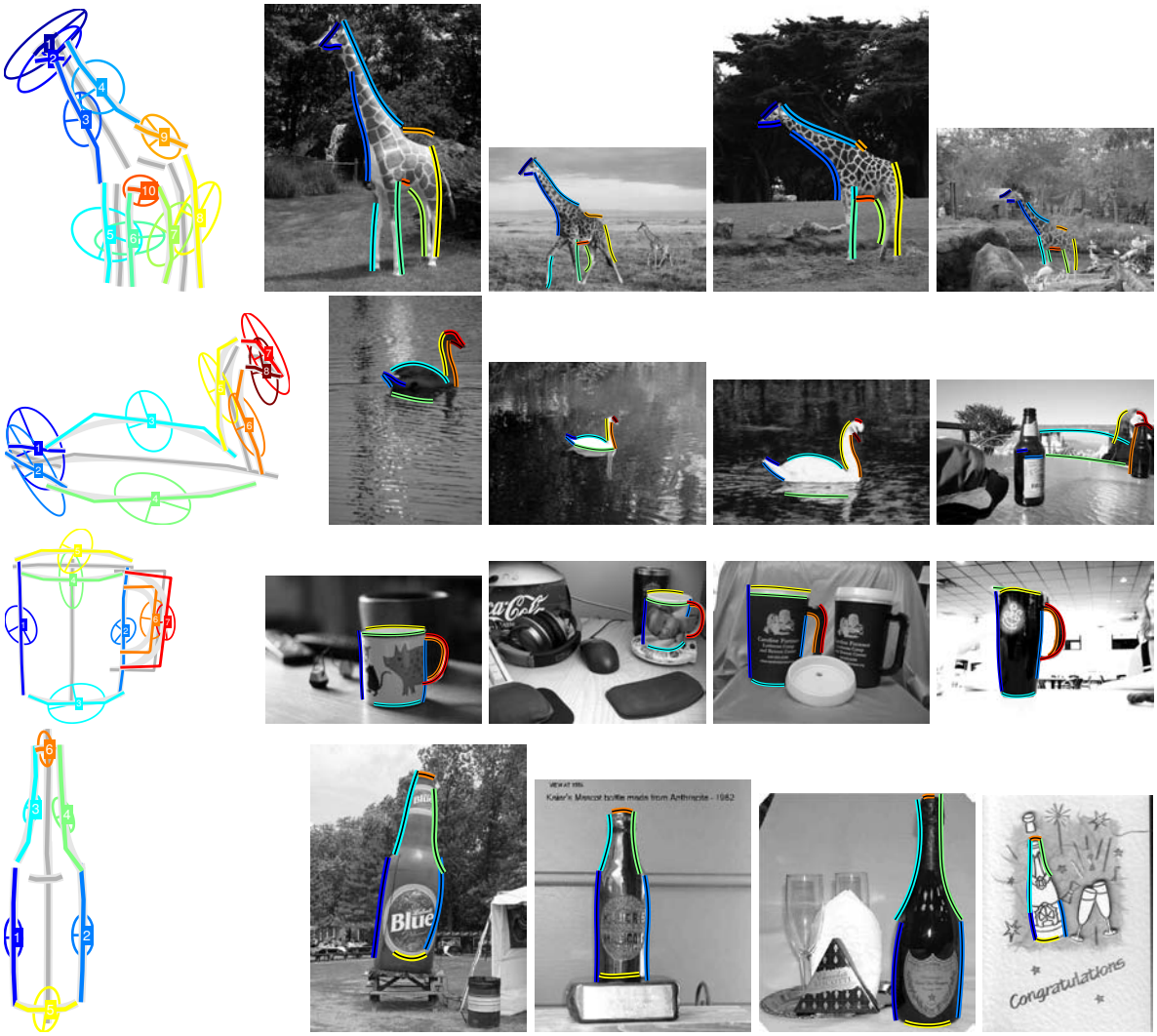


Figure 5.3: Learned ETHZ Shape Classes models (left) and example detections (right). For models, mean local part shapes and selected mean symmetry axes are drawn at mean positions and relative scales. Covariances of part positions are shown as ellipses.

5.4 KNOWLEDGE TRANSFER

In the following, we highlight two different levels of inter-model knowledge transfer supported by our model. First, we show that a full model, learned for a known class A, can be readily transferred to a new but related object class B. An object model for this new class B is obtained from one or a few training instances plus the transferred knowledge from the previously learned object class A. Second, we also show *partial knowledge transfer* by restricting a previously learned model A to a proper subset of parts, retaining all knowledge about their spatial configuration, relative scales, and symmetry relations. The resulting partial model can be transferred to a new class B for which only a few training instances are available.

5.4.1 Full model transfer

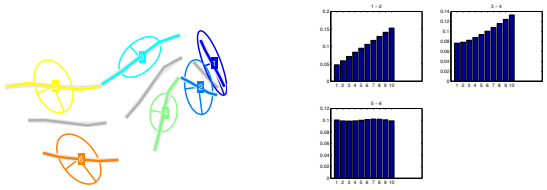
Our approach to combine prior knowledge and data is inspired by, but not strictly adhering to, the Bayesian paradigm. Instead of deriving a posterior distribution over models, given data D, from a prior and corresponding likelihood $p(\theta|D) \propto p(\theta) \times p(D|\theta)$, we follow the simpler route of directly combining and manipulating components of models that we have learned. These manipulations are valid because of the specific factorization and parametric forms of involved distributions. In particular, since all distributions are Gaussian, we can manipulate means and covariances separately, and can restrict models to subsets of parts by marginalizing out the ones we are not interested in.

Let $m_A(\mu_A, \Sigma_A)$ and $m_B(\mu_B, \Sigma_B)$ be two models, where m_A is the *base model*, i.e., the model which we want to transfer, and m_B a model learned from k training instances of class B. We denote m_B a k-shot model. Now the question arises which knowledge should be transferred from m_A to obtain a more powerful model for class B. Consider, e.g, the case that class A corresponds to *horses* and class B corresponds to *giraffes*. While the mean of the overall object shape is different the variation in object shape is similar as both classes belong to the class *quadrupeds*. Therefore we derive a combined model $m_{AB}(\mu_{AB}, \Sigma_{AB})$ for class B by taking μ_{AB} to be μ_B, and Σ_{AB} to be a weighted combination of Σ_A and Σ_B. For $k = 1$, we set $\Sigma_{AB} = \Sigma_A$. The experiments in Sect. 5.5.1 show results of this procedure.

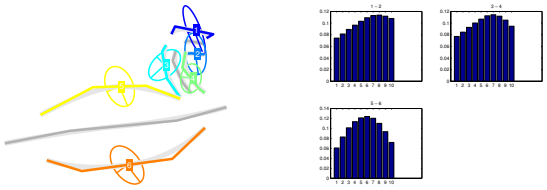
5.4.2 Partial model transfer

The factorization of our model into separate components for local part shape, relative scales, symmetry relations, and the overall spatial layout facilitates keeping subsets of parts, while discarding others. For part shape as well as relative scale components, we keep all relevant part contributions. For symmetry relations, we keep all contributions involving at least two relevant parts. For spatial layout, we can marginalize out all irrelevant parts.

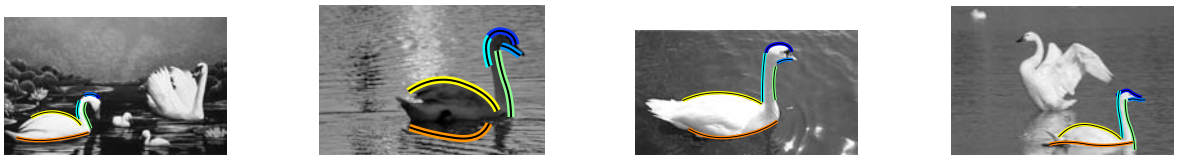
To realize the importance of *partial knowledge transfer* consider the following



(a) Partial horse model



(b) Combined swan model



(c) Example detections using the combined swan model.

Figure 5.4: Partial transfer models (a)(b), and *1*-shot detections (c).

example (see also Sect. 5.5.2). Let us assume class A again corresponds to *horses* and class B corresponds to *swans*. As the first is a quadruped and the second is a bird, one might see little opportunity for knowledge transfer, since global object shape is different.

However, there is indeed *partial* knowledge that can be transferred, namely the topology of a subset of parts (head, neck, and torso). As this information is contained in the *horse* model, we may first extract the corresponding relevant portion by marginalization, and then transfer this partial knowledge. The experimental section shows the usefulness of such partial knowledge transfer, which we argue to be a very general and versatile concept, as many parts and constellations of parts reoccur across many object classes. Therefore, use of such *partial knowledge transfer* about constellation, local shape and symmetries of object parts and part ensembles is a powerful tool to enable scalability to large numbers of object classes.

5.5 KNOWLEDGE TRANSFER EXPERIMENTS

We demonstrate the ability of our approach to effectively transfer knowledge between models by a series of recognition experiments based on the animal quadruped classes *horse*, *elephant*, *cheetah*, and *giraffe* for which we combined images from the Mammal Images Benchmark (Fink and Ullman, 2008), the Corel data base, INRIA Horses (Jurie and Schmid, 2004), and additional images from the web. Images show quadrupeds roughly pose-aligned, but at varying scales, and contain considerable background clutter (see Fig. 5.1). While all quadrupeds share a common topology (head, neck, torso, and four legs), they vary significantly in the concrete embodiment, leading to variations in both the appearance of individual body parts as well as their spatial layout. In addition, we use the *swan*, *mug*, and *bottle* classes from the ETHZ Shape Classes data set in Sect. 5.5.2 for partial knowledge transfer.

All experiments follow this protocol: Models are learned from a set of training

images of a given class and evaluated on a test set consisting of images containing at least one instance of that class, and a comparable number of background images not containing any class instances. Performance is evaluated in a retrieval setting where we run detection for each test image and record the highest scoring hypothesis. For each n between 1 and the number of test images, we plot the fraction of images belonging to the class in the n highest scoring ones.

5.5.1 Full model transfer

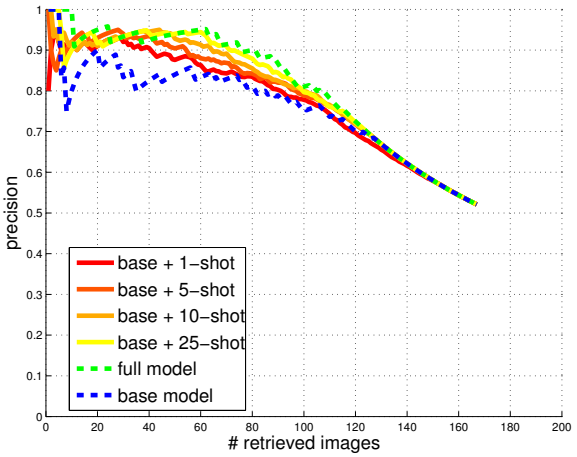
Using the quadruped classes, we show that prior knowledge about the general stature of a quadruped can be used to bootstrap specialized quadruped detectors. In particular, we learn a *base model* from all 170 positive INRIA *horse* images, which we assume to yield a reasonable model of quadruped stature (see Fig. 5.7). We then use k training images of another quadruped class and learn a k-shot model from these images. The models are combined as described in Sect. 5.4 and the combined model is evaluated as above. We found experimentally that the weighting of the individual models has little impact on performance and thus report all results for uniform weighting.

Fig. 5.5 gives recognition results for the classes *elephant*, *cheetah*, and *giraffe* without and with symmetry relations. Each plot compares the performance of combinations of the base model with k-shot models learned from $k \in \{1, 5, 10, 25\}$ training images, the *full model* learned from all available training images of the target class, and the base model alone. The curves for $k \in \{1, 5, 10\}$ are averaged over 5 different random choices of k training images among the full 25 training images available for each class.

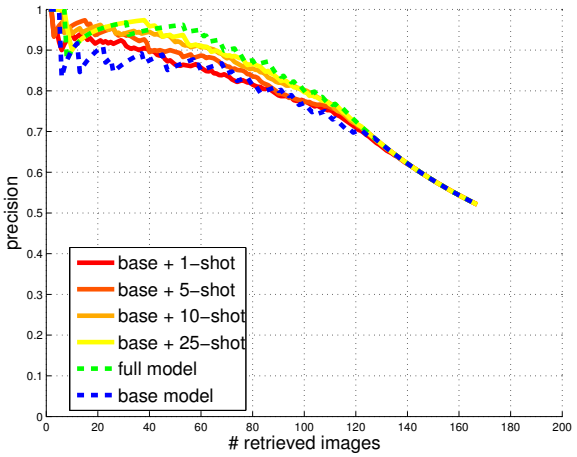
We first observe in Fig. 5.5 that the base model learned entirely from *horse* images performs surprisingly well on *elephants* and *cheetahs* despite major differences in appearance. It can therefore be transfered directly even without a single training image. This can be explained by the fact that the *horse* model already captures a fair amount of variations in the shape and spatial layout of *elephants* and *cheetahs*. This is also confirmed in Fig. 5.7 and 5.8: means and covariances of part shape as well as constellations of full *elephant* and *cheetah* models are visually close to the horse base model (Fig. 5.7). Furthermore, all shown symmetry distance models share common properties, namely, an almost linear increase in distance between head parts (1-2), quadratic dependency between pairs of leg parts (5-6 and 7-8), and the almost flat shape of the torso (9-10).

Adding training images clearly improves precision and adapts models to the target classes. A small number of training images (5 for *cheetah* and 10 for *elephants*) is sufficient to achieve a performance that is largely equivalent to the corresponding full model. Fig. 5.8 confirms this observation: combinations of 5-shot and base models (middle column) are visually close to the corresponding full models (right column) and can thus be expected to behave comparably.

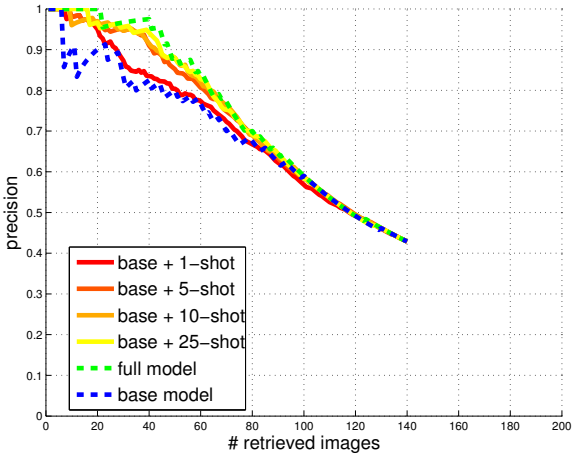
Interestingly, the base model performs poorly for the *Giraffe* class as the full *giraffe* model differs quite strongly from the *horse* base model (e.g., the neck parts, see Fig.



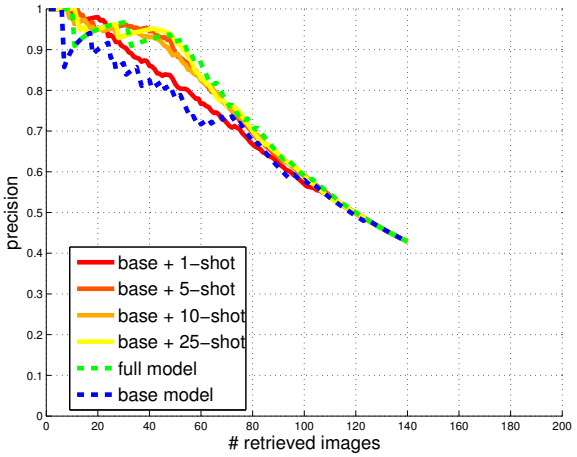
(a) Elephant precisions



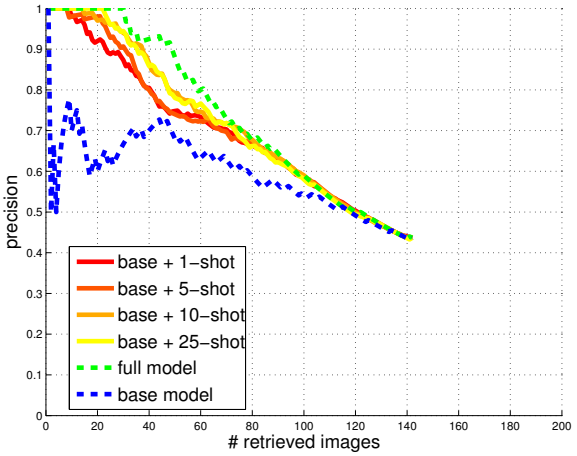
(b) Elephant precisions, SLS



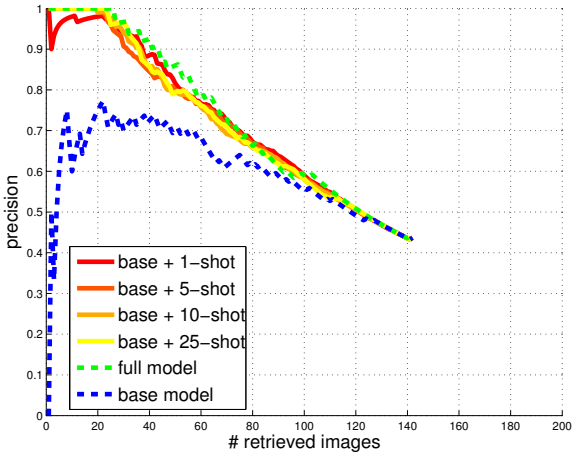
(c) Cheetah precisions



(d) Cheetah precisions, SLS



(e) Giraffe precisions



(f) Giraffe precisions, SLS

Figure 5.5: Full model transfer recognition results without symmetry relations (left) and with symmetry relations (SLS, right).

5.8(c)). Note however, that even a single training image is sufficient to boost the performance to almost the level of the full model. This is particularly pronounced for added symmetry relations, and explained by the high degree of similarity among all symmetry distance models.

In order to understand the role of the base model, we further compared the performance of combined *1*-shot models with *1*-shot models using isotropic regularization, which we determined empirically on a separate data set (ETHZ Shape). Even though these models can perform on a similar scale as the combined models, they tend to be slightly worse on average, and introduce the disadvantage of having to choose a suitable regularizer, while regularization comes for free with a base model.

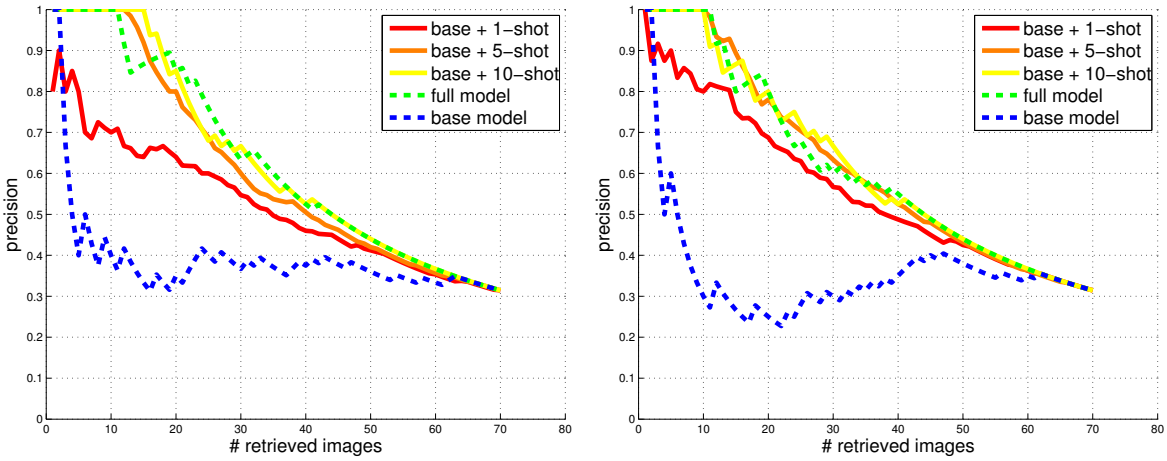
To summarize, knowledge transfer with a suitable base model clearly reduces the number of required training images in all cases. *k*-shot models including symmetry relations between parts are often superior. Also, the variance of the curves including symmetry relations exhibit less variation, in particular for *giraffes* (Fig. 5.5(f)) clearly showing the importance of symmetry relations for knowledge transfer.

5.5.2 Partial model transfer

For partial model transfer, we restrict a base model to a proper subset of parts and combine this restricted base model with a *k*-shot model of a new class. As mentioned before we can transfer partial knowledge of a *horse* base model to the *swan* class. Therefore, in a first experiment, we restrict a *horse* model to head, neck, and torso parts and then combine this restricted base model with *k*-shot models of *swans* (see Fig. 5.4). In the second experiment we transfer partial knowledge of a *mug* base model to the *bottle* class. For this we restrict the *mug* model to the sidewall and bottom parts, discarding handle and upper rim parts and combine this with *k*-shot models of *bottles*. As before we report retrieval performance for *swan* and *bottle* images respectively.

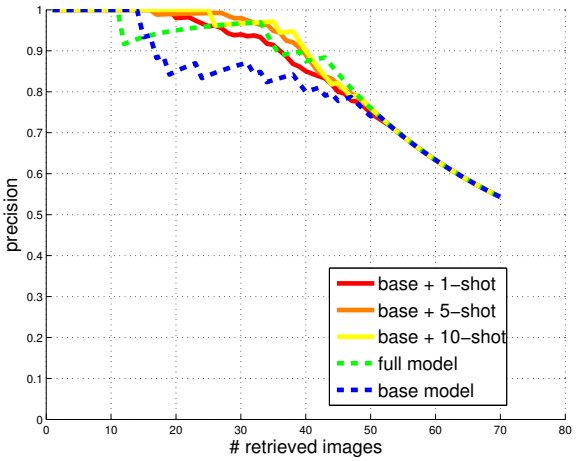
From Fig. 5.6(a) and (b), it is immediately apparent that the restricted *horse* base model performs only at chance level for the *swan* retrieval, both with and without symmetries. Strikingly, adding a *single image* of a *swan* drastically improves detection rate (base + *1*-shot). As before, adding only a handful of images to the restricted base model yields performance close to the full model. Likewise, adding symmetries to the model is highly beneficial. In particular, the combined *swan 1*-shot model benefits significantly ($\approx 10\%$) from including symmetry relations.

In the second experiment (Fig. 5.6(c) and (d)) the *mug* base model already enables to retrieve *bottle* images quite well. This is due to the fact that the two classes not only share several common parts, but their shape is also similar. In this case, already a single training example is sufficient to reach the performance level of the corresponding full models. From these experiments we can conclude that our model does indeed allow for partial knowledge transfer and enables to train object models from few training images. In cases where object classes share many similarities (*mug-bottle*-transfer) as little as one training instance can suffice. For larger variations between objects (*horse-swan*-transfer) five training instances can yield a good model.

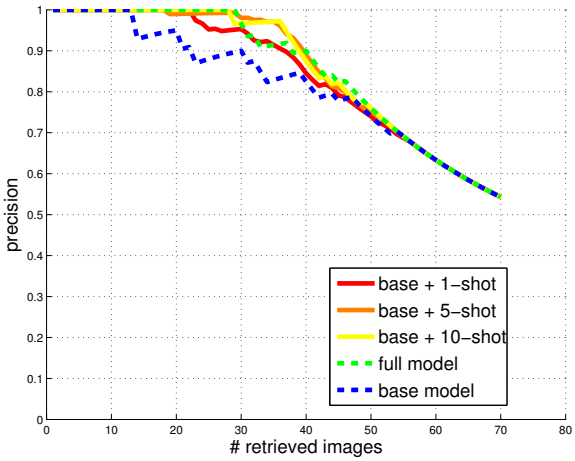


(a) Swan precisions

(b) Swan precisions, SLS



(c) Bottle precisions

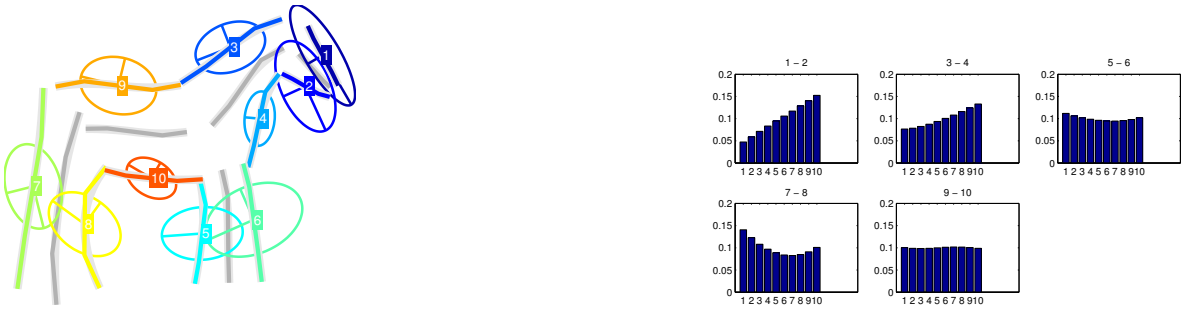


(d) Bottle precisions, SLS

Figure 5.6: Partial model transfer recognition results without symmetry relations (left) and with symmetry relations (SLS, right).

5.6 CONCLUSIONS AND FUTURE WORK

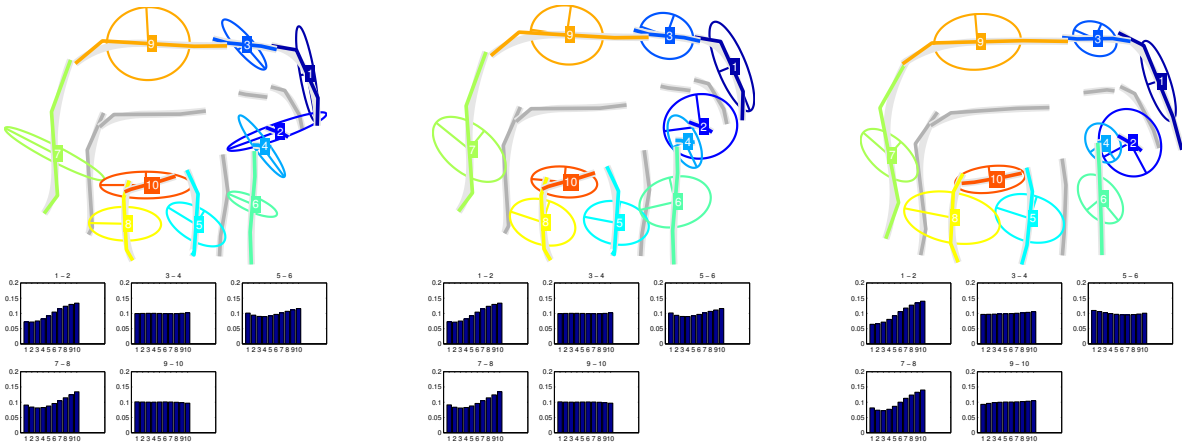
While pioneering work on knowledge transfer for object class model training exists, none of it has been adopted widely. Despite this fact, we strongly believe that knowledge transfer is an important ingredient to enable learning and recognition of large numbers of object classes. As demonstrated by our results, our shape-based model enables explicit knowledge transfer between object classes thereby simplifying training for new object classes. The model's ability to transfer individual components makes our approach applicable to a large number of scenarios. Its competitive results on the ETHZ Shape Classes confirm the validity of the object model formulation for object class detection. The use of local symmetries improves the performance both for detection and model transfer significantly although symmetries are so far seldom used for object detection. Since both, the model as well as the proposed DDMCMC inference method, can be easily extended to larger number of parts and to include other complementary features, we believe that it presents large opportunities for future work.



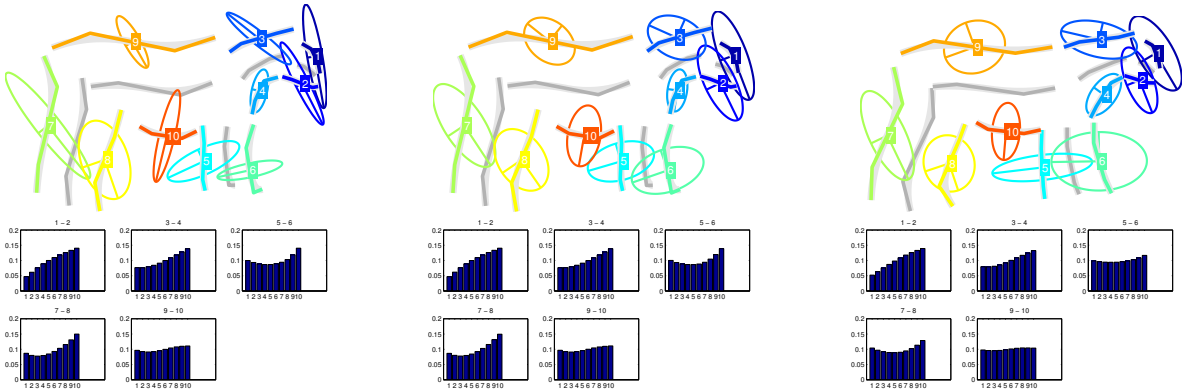
(a) Base model

(b) SLS symmetry line length profiles

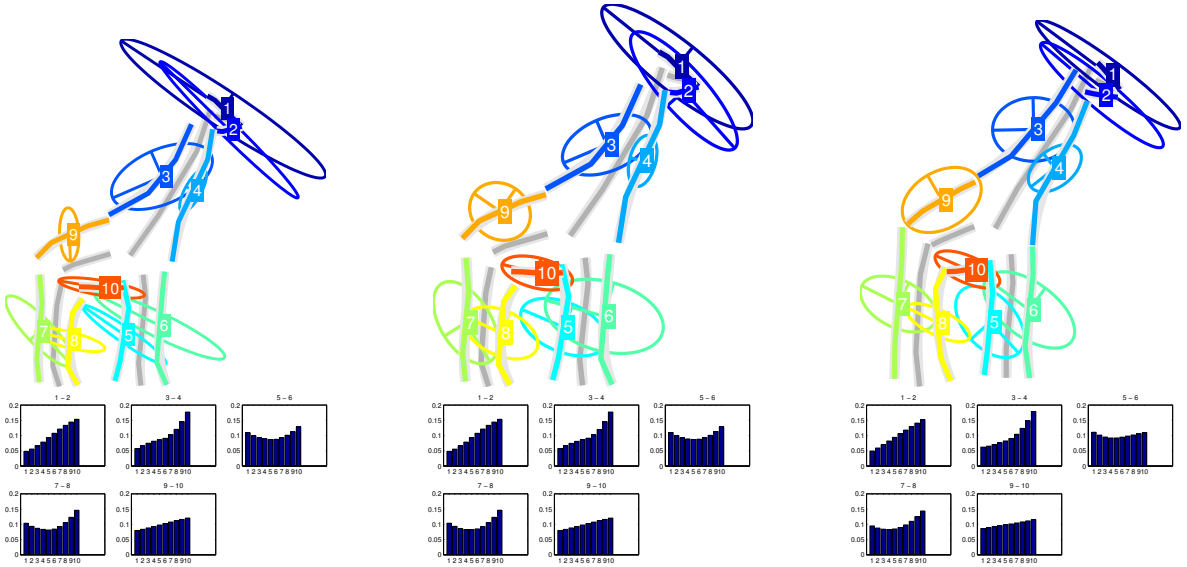
Figure 5.7: The horse base model used in the *k*-shot experiments of Sect. 5.5. For clarity, we show only a subset of symmetry relations: numbers above plots in (b) refer to pairs of part numbers in (a).



(a) Elephant



(b) Cheetah



(c) Giraffe

Figure 5.8: Animal models: left: 5-shot model, middle: 5-shot model combined with base model, right: full model. Note the similarities between the models on the right and in the middle.

SHADING CUES FOR OBJECT CLASS DETECTION 6

Contents

TWO-DIMENSIONAL SHAPE has been the focus of the previous chapters, either on the level of individual local features (Chapter 3), groups of local features (Chapter 4), or in combination with their spatial layout and pair-wise symmetry relations (Chapter 5). This chapter focuses on visual hints on three-dimensional shape, arising from shading artifacts on curved surfaces, for the specific case of cylindrical surface primitives. In analogy to two-dimensional shape, we expect these shading cues to be generic in nature, and thus potentially transferable between object classes, although the presented experiments are limited to a single object class. The applicability of the proposed shading model to real world images of a standard benchmark data set is verified in combination with the shape-based object class model presented in Chapter 5. In particular, the shading model is used to verify detection hypotheses delivered by the shape-based object class model.

6.1 INTRODUCTION

In recent years, impressive progress has been reported in the recognition of a wide variety of object classes. Object models based on robust local appearance features (Lowe, 2004), in combination with bag-of-words (Csurka *et al.*, 2004), or more spatially constrained models (Fergus *et al.*, 2003) perform well on recognition benchmarks. More recently, 2D-shape-based approaches have also shown to yield comparable performance (Ferrari *et al.*, 2007).

Interestingly, none of these 'modern' recognition approaches makes explicit use of 3D shape information provided by shading cues. This is in contrast to early approaches in object class recognition and also contrary to intuition, since humans make extensive use of shading information to assess object shape (Kleffner and Ramachandran, 1992; Koenderink *et al.*, 1996), which is important for recognition.

Figure 6.1: Shape-based object detections and shading cues on *ETHZ Mugs*. From left to right: (1) original image, (2) shape-based detection, (3) shading cue based on (2), back-projected into the image. The green arrow reflects estimated lighting direction, seen from above the scene.

One might argue that at least part of the shading information is encoded implicitly by appearance features, and thus available to 'modern' recognition algorithms. This comes at a cost, however: in order to reliably separate possibly relevant shading information from background, these algorithms need to use statistics over large numbers of training samples. Explicitly modeling or learning shading information can remedy this problem, by encoding relevant information into the model itself.

Inferring the shape of a surface from shading is unfortunately a difficult problem, and has long been a major focus of computer vision research. By nature, shape-from-shading (SFS) is highly ambiguous: without any prior knowledge, a given image of an observed scene could have been generated by an infinite number of different combinations of object surfaces present in the scene, their reflectance, and lighting conditions. As a consequence, SFS approaches are typically restricted to controlled environments, or reduce ambiguity by imposing strong assumptions on surface shapes, material, and lighting (Horn and Brooks, 1989).

As a consequence, making direct use of SFS for object recognition in natural images has proven difficult although there have been many attempts. Worthington and Hancock (2001) apply Shape-from-Shading techniques to object recognition, on the level of individual object instances (COIL-20 data set (Nene *et al.*, 1996)). Their work builds upon a mid-level representation of surface topography based on local curvature and shape-index (Koenderink and van Doorn, 1992) information, and uses histograms and region descriptors on top of this representation. Following a similar route, Lichtenauer *et al.* (2005) suggest using orientation and curvature of isophotes (lines of equal brightness) as features in a classification framework for classifying image patches as face/non-face. Wu *et al.* (2007) report improved performance for gender-classification of pose-aligned face images with needle-map features obtained via shape-from-shading. Nillius *et al.* (2008) present generic shape detectors for cylinders and spheres, using model-based PCA, and a multi-scale sliding-window search over image regions. Mori *et al.* (2004) describe shading on human limbs by prototypical, half-wave rectified gradient image patches, and use a similarity score in order to identify candidate limb image regions.

While these relatively recent approaches use SFS as bottom-up features, more than ten years ago, Haddon and Haddon and Forsyth (1998) suggested a promising alternative, by verifying given 3D shape hypotheses in a top-down fashion using shading cues. In line with Biedermann's theory of recognition-by-components (Biedermann, 1987), and similar in spirit to Weinshall (1992), the authors suggest shading primitives as the basis for recognition. The recognition problem amounts to finding valid configurations of several primitives.

Borrowing from these ideas, we use a part-based object class model at the core of our approach. We explicitly model the 2D shape of individual parts, together with pairwise, semi-local symmetry relations, and the overall spatial layout. We then establish 3D shape hypotheses based on object parts and shading cues, and add them as additional cues to the final detection hypothesis. In particular, this chapter makes the following contributions:

- We propose a shading model for cylindrical surface primitives, which we show to yield acceptable model fits on real world images, taken from a standard object detection benchmark (Ferrari *et al.*, 2006b), and analyze the failure cases.

- We present first results to integrate this shading model as an additional cue into an existing state-of-the-art shape-based object detection framework.

- We give quantitative experimental evidence that shading cues can indeed increase recognition performance.

The remainder of this chapter is organized as follows. Section 6.2 introduces the shading model. Section 6.3 reviews the shape-based object detector. Section 6.4 gives experimental results, and Section 6.5 concludes with an outlook on future work.

6.2 SHADING MODEL

Similar to the work of Haddon and Forsyth (1998), our shading model follows the principle of hypothesis verification. Instead of recovering the 3D shape underlying an image area in a bottom-up fashion, it starts from a given 3D shape hypothesis and tries to verify this hypothesis based on image evidence. In particular, the observed image evidence must be consistent with the 3D shape, some estimated reflectance properties, and the estimated scene illumination. Proper regularization is required since the estimation is highly ambiguous — the same image can be generated by different combinations of surface shape, reflectance, and lighting.

6.2.1 A shading primitive

In the following, we present a concrete instantiation of this hypothesis verification framework for the case of cylindrical surfaces. Our model starts with the hypothesized occluding contours of a cylindrical shape (the cylinder side-walls) in the image plane and tries to verify this hypothesis based on evidence from the pixels on the cylinder surface using a simple model for lighting and reflectance. (Figure 6.1 shows some successful examples on images from the *ETHZ Mugs* dataset.)

We assume that the directional lighting in the scene can be well approximated by a single point light source located far away from the surface of interest. In the limit, i.e., for infinite distance this corresponds to a directional light source. We model the remaining contribution as ambient illumination impinging on the surface uniformly from all directions. Both components of the model can be simply added due to the principle of superposition.

Regarding reflectance, we restrict ourselves to the simplest possible model and assume that the surface is diffuse (Lambertian) with a constant albedo (Dorsey *et al.*, 2007). Specular effects of surface texture are ignored and will be treated as outliers during parameter estimation. This model implies that barring occlusion effects the reflected radiance depends solely on the direction of incident radiance relative to the surface normal. All points with equal surface normals will exhibit equal brightness in the image.

Shading on cylindrical surfaces. Let us assume an orthographic projection of a cylindrical surface, with the viewing direction being perpendicular to the cylinder axis. We divide the surface into a set of circular cross-sections, such that the viewing direction is parallel to the corresponding sectional planes. A point on the observed half of a cross-section can then be described by the parameterization ϕ (see Figure 6.2). Due to orthographic projection, $s = \sin\phi$ can be used to parameterize the projection of the cross-section onto the image plane without introducing any distortions. We can now establish a functional dependency between s and the observed image values for the corresponding surface point $B(s)$. Note that we need to ensure that the image is in photometrically linear space. This typically requires applying an inverse

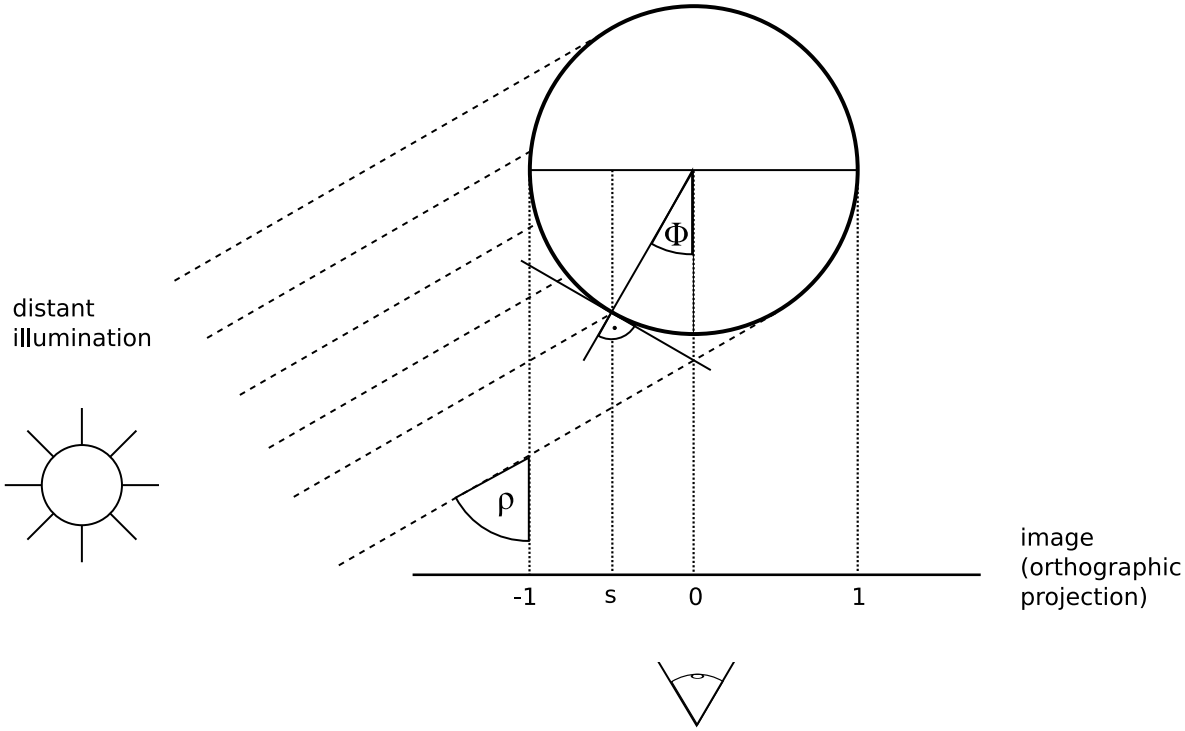


Figure 6.2: Geometry of visible cylinder (half) cross section, parameterized by $s = sin\phi$, viewing direction, and light source at infinity. ρ denotes the angle between illumination direction and (orthographic) viewing direction.

gamma correction. Let ρ be the angle between the direction of incident light and the viewing direction, both projected on a plane perpendicular to the cylinder axis. The observed image value is then

$$B(s) = a + b * \max(0, \cos(\rho - \phi(s))). \tag{6.1}$$

The two scaling factors $a > 0$ and $b > 0$ determine the intensity of the ambient and the directional lighting, respectively, multiplied with the albedo. The maximum in Equation 6.1 ensures that surface points with normals pointing away from the directional light source, and which are therefore in shadow, do not contribute physically invalid, negative radiance.

Let us now assume viewing the cylinder from an elevated angle and/or rotating the camera around the viewing direction. The corresponding cross-sections are no longer perpendicular to the cylinder axis, and change their shape from circular to elliptical. As a consequence of both orthographic projection and directional lighting, these elliptical cross-sections can be transferred into equivalent circular cross-sections by sliding all constituent points along the cylinder's isophotes, parallel to the cylinder axis. Following this argumentation, Equation 6.1 can be proven valid for any cylinder cross-section without changing the parameterization s, as long as its projection on the image plane is a straight line connecting the two cylinder side-walls.

Implementation. In order to determine the model parameters a, b, and ρ we need to select a set of cross-section points

$$\{s_i\}_{i=1}^n, s_i \in [-1, 1] \tag{6.2}$$

and corresponding brightness values

$$\{b_i\}_{i=1}^n, b_i \in [0, 1]. \tag{6.3}$$

We obtain pairs of the form (s_i, b_i) by first sampling a fixed number of equidistant points on the two occluding contours of a hypothesized cylindrical surface and then connecting corresponding pairs of points by straight lines. We finally sample pixel brightness values b_i by parameterizing each line by $s_i \in [-1, 1]$ using the Bresenham algorithm (Bresenham, 1965).

The parameters a, b, and ρ can now be determined using standard non-linear least squares optimization techniques, such as the Levenberg-Marquardt algorithm (Levenberg, 1944), by minimizing the sum of squared residuals

$$S(a, b, \rho) = \sum_{i=1}^{n} (b_i - B(s_i))^2. \tag{6.4}$$

In practice, we observe that the non-differentiable $\max(.)$ function does not pose any problems during optimization.

Since surface texture, specular reflections, and other unmodeled effects often yield a significant number of outliers (see, e.g., the textured mug in Figure 6.3(a)), we use RANSAC (Fischler and Bolles, 1981) to select a single consistent model. Invariance w.r.t. global brightness variations is achieved by selecting inliers according to a threshold on the squared residual $(\log b_i - \log B(s_i))^2$ in logarithmic space.

6.2.2 Example shading model fits

In order to demonstrate the validity of the proposed cylindrical shading model, we give qualitative as well as preliminary quantitative results on the *Mug* category of the ETHZ Shape Classes dataset (Ferrari *et al.*, 2006b). Figures 6.3 and 6.4visualize exemplary shading model fits of varying quality on eight different images, starting from shape-based object detections (see Section 6.3). In particular, we select the single best true positive *Mug*-hypothesis per image, each consisting of seven parts (left and right side-walls, upper and lower rims, bottom, and two handle parts), and fit a cylindrical shading model between the side-walls of the *Mug*.

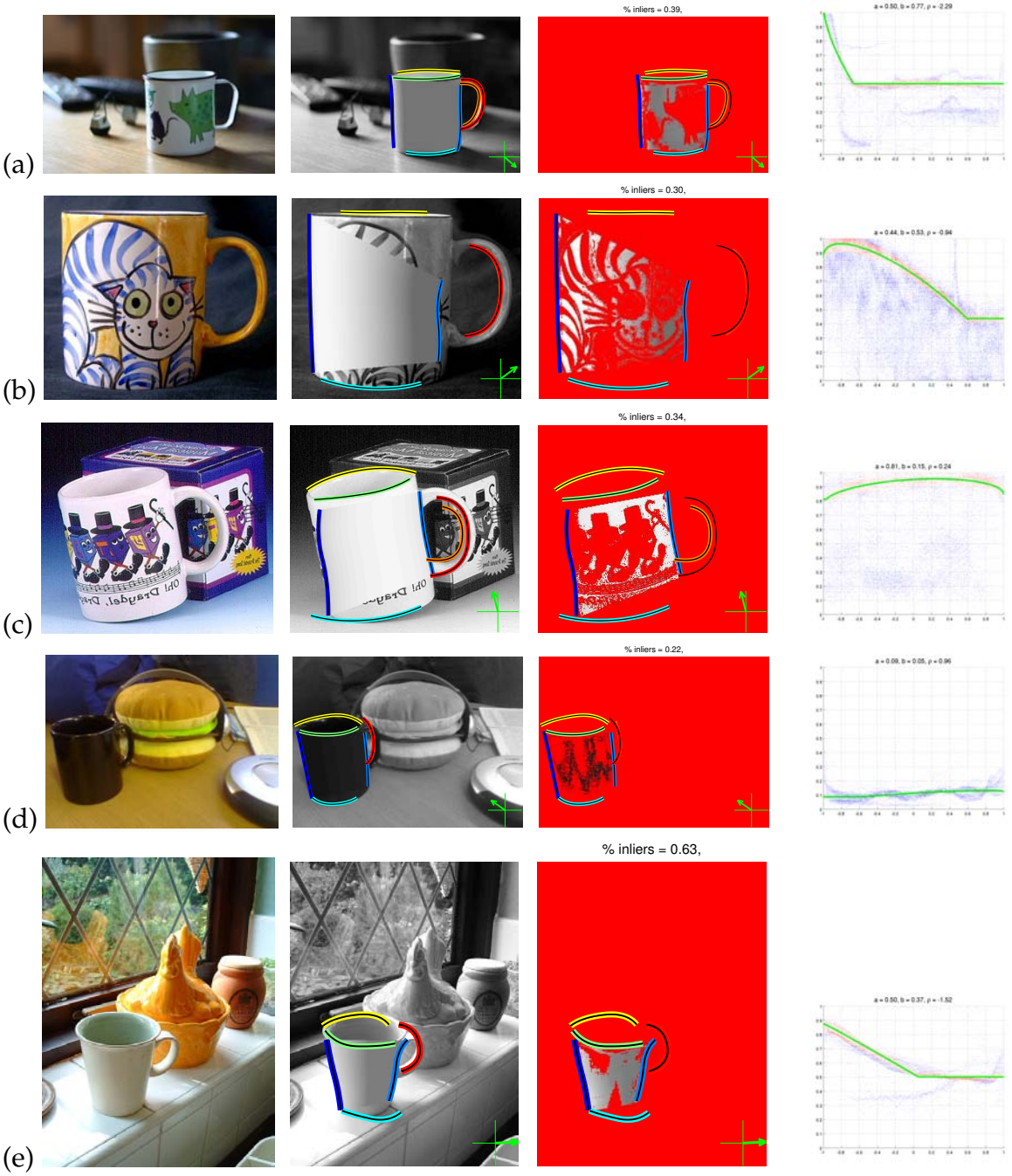
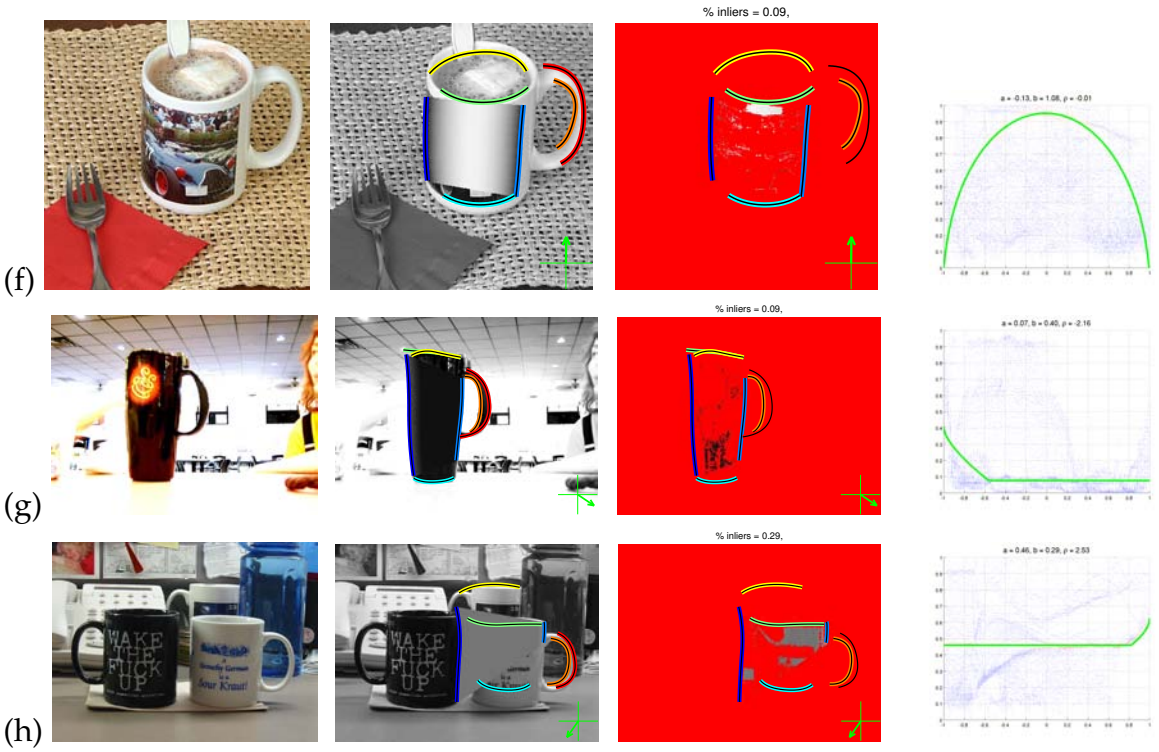


Figure 6.3: Example shading fits, based on shape-based object detection hypotheses (see Section 6.3). First column: original image. Second column: back-projected shading model. Third column: RANSAC inliers and estimated lighting direction, seen from above the scene. Fourth column: shading model fit with accepted samples (red) and outliers (blue). *Near perfect* fits: (a) - (d), *acceptable* fit: (e).



(f) (g) (h)

Figure 6.4: Example shading fits, based on shape-based object detection hypotheses (see Section 6.3). First column: original image. Second column: back-projected shading model. Third column: RANSAC inliers and estimated lighting direction, seen from above the scene. Fourth column: shading model fit with accepted samples (red) and outliers (blue). *Failures* ((a) - (c)) due to object texture (a), specularities (b), erroneous shape fits (c).

Fit quality	Shape	Shading on GT	Shading on shape fit
Failure	0.14	0.06	0.27
Acceptable	0.18	0.25	0.16
Near perfect	0.68	0.69	0.57
Non-failure	**0.86**	**0.94**	**0.73**

Table 6.1: Quality of shape and shading model fits. For shading model fit quality, we distinguish between ground truth (GT) and shape model fit as the basis for computing the shading model. The last row summarizes the two preceding rows.

Failure case	Specularities	Shape fit	Lighting	Texture
Fraction	0.42	0.32	0.16	0.11

Table 6.2: Likely failure reasons of shading model fits.

6.2.3 Discussion

Table 6.1 gives an assessment of the quality of obtained shading fits on *Mug* objects. It compares shading fits obtained by using detection hypotheses from the shape-based object detector with fits obtained from ground truth annotations of *Mug* side walls. Since we obtain ground truth side walls by marking actual shape features in images, these annotations are available only for 36 of 44 *Mug* images, due to imperfect shape feature detections. The table further gives estimates on the quality of the original shape fits, as these constitute the basis for shading fits. Since ground truth on the lighting conditions in which the images were taken is hard to acquire in retrospect, shading fit quality is assessed by visual inspection, and roughly categorized into *near perfect* (all parameters sensibly fit), *acceptable* (parameter estimates deviate slightly from human assessment), and *failure* (clearly erroneous parameter estimates).

We note the following observations: First, in 0.94 of the cases, an at least *acceptable* shading model can be fit from available ground truth occluding contours. This indicates that the proposed shading model is in principle capable of modeling most shading artifacts present on the tested ETHZ *Mug* images, despite variations in shape, texture, material, and lighting. Second, despite the fact that this number decreases significantly if shape model fits are used as a basis an encouraging amount of 0.73 of the obtained shading models is still at least *acceptable*. These models correctly reflect the cylindrical 3D shape of the *Mug* objects, and can thus be beneficial for recognition.

Table 6.2 lists the most likely reasons for imperfect fits, again assessed by visual inspection. The most frequent likely reason (0.42) for failure is the presence of specularities and reflections, which are not explicitly included in the shading model, but possibly rejected as outliers by RANSAC. Figure 6.4(b) gives an example of an erroneous fit, caused by the highly specular, dark material of the mug.

The second most frequent reason for failure is the sometimes insufficient quality of shape model fits used as the basis for shading. According to Table 6.1, 0.14 of these

shape fits are failures, resulting in erroneous support for the shading model. Figure 6.4(c) shows an example, where pixels on the mug and pixels from the background are wrongfully combined in the set of selected inliers.

Figure 6.3(e) gives an example of a still acceptable fit, showing a deviation in the estimated incident light direction from what one would expect: contrary to intuition, the incident light is estimated as coming strictly from the left, and not from the direction of the window. This is an instance of difficult lighting conditions, and attributed to 0.16 of the failure cases.

Surprisingly, texture is rarely a source of confusion (0.11 of the cases). Figure 6.4(a) shows one of the few examples where object texture (a photo printed onto the mug) is wrongfully picked up by the shading model (the shape fit for this example is also imperfect; the corresponding shading fit for ground truth side-walls is in fact *near perfect*). Figure 6.3(a) - (c) gives examples of successfully handled textures.

6.3 SHAPE MODEL

Our approach to integrate shading cues into object recognition is based on the shape-based object class detector presented in Chapter 5. In particular, we use information from this model in two different ways: 1) the shading model described in Section 6.2 is used to verify hypotheses provided by the shape-based part-detections (cylinder side walls), and 2) a final score is calculated by combining the shape-based detection scores with the fitted shading model parameters (detailed in Section 6.4).

6.4 EXPERIMENTS

The following examines the potential benefit of adding our shading cue for object recognition. To integrate our shading cue into the probabilistic model of the shape-based object detector described in Chapter 5, we combine the outputs of both models in a discriminative framework (sometimes referred to in the literature as *late integration*). In particular, we train two linear SVM classifiers. The first is using the shading model parameters a, b, ρ, the fraction of inliers, and the mean squared residual on the inliers. The second additionally uses the shape-based detection score.

As in Section 6.2, we base our evaluation on the category *Mug* from the ETHZ Shape Classes data set (Ferrari *et al.*, 2006b). We set up a binary classification task as follows: for each of the 251 images (44 *Mugs*, 207 non-*Mugs*) of the data set, we select the highest scoring detection hypotheses for the category *Mug*. We then either store it as a positive (in case it qualifies as a true positive detection according to an overlap criterion) or as a negative (in case it does not) training example. We then train and test classifiers on these examples using 5-fold cross validation, in order to have a reasonable amount of positive training examples available. Each model is individually optimized w.r.t. the maximum margin-training error minimization tradeoff parameter C of the linear SVM. Please note that this experiment is different from the original setup in (Ferrari *et al.*, 2007) and therefore does not allow for direct

comparison. However, as a first proof of concept and to understand the potential benefit of our shading cues for recognition we consider this experiment appropriate for the purpose of this chapter.

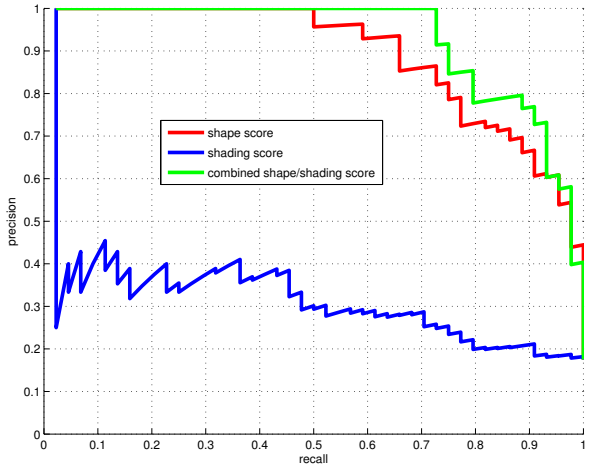


Figure 6.5: Precision/recall curves for classifying shape-based detection hypotheses into *Mugs*/non-*Mugs*, based on different scores.

Figure 6.5 plots precision and recall curves for the binary classification task described above. While the red curve is based on the shape-based detection score alone, the blue and green curves have been obtained by varying a threshold on the corresponding SVM scores, obtained by 5-fold cross validation.

Observations. First, the classifier trained on shading model parameters alone essentially fails to discriminate between positive and negative examples. This is not surprising, since it fully neglects valuable information about the shape and spatial layout of *Mugs*, while concentrating fully on cylindrical shading (as, e.g., in Figure 6.6(i),(j)).

Second, the shape-based detection score shows good performance (Equal-Error-Rate (EER) of 77.3%), despite the negative examples being hard ones (since we picked the highest-scoring ones per image).

Third, combining the shape-based detection score and the shading model parameters yields a considerable improvement over the shape-based detection score. The shading cue improves recall at precision level 100% from 50% to 72.7%, and lifts EER from 77.3% to 79.6%. Figure 6.6(a) - (h) depicts complementary detections hit and missed by the two classifiers, respectively. It lists high scoring detections correctly classified by one, but mis-classified by the other classifier, at the highest achieved recall for precision 1.0. Apparently, the combined shape-shading classifier makes efficient use of available shading information, compensating weak shape model fits (Figure 6.6(c) - (f)). The two examples mis-classified by the shape-shading combination can be attributed to imperfect shading fits due to specularities and

texture, respectively. Figure 6.6(i),(j) show two false positive classifications of the combined shape/shading score. While the bottle label is in fact an instance of cylindrical shading, the water surface underneath the swan is clearly an error.

6.5 CONCLUSIONS AND FUTURE WORK

In this chapter, we have introduced a shading model for cylindrical surface primitives, based on hypothesis verification, and demonstrated its validity on images of a standard data set for shape-based object detection. We have shown preliminary results of integrating this shading model as an additional cue into an existing, state-of-the-art, shape-based object detection framework, and obtained quantitative experimental evidence for its potential usefulness in recognition.

Based on these encouraging results, we consider the proper integration of the proposed shading cue into the Data-Driven Markov Chain Monte Carlo framework of Chapter 5 an obvious next step, as well as adding more 3D surface primitives, such as spheres and corners.

Figure 6.6: (a) - (h): complementary detections. (a) - (f): six high scoring *Mug*-hypotheses correctly classified by the combined shape/shading score, but missed by the pure shape score (precision level 1.0, highest recall). (g),(h) the only two hypotheses for the inverse case. (i), (j) Two false positives of the combined shape/shading score at EER.

7 LEARNING SHAPE MODELS FROM 3D CAD DATA

Contents

Exploiting additional sources of knowledge as an alternative to real world training images is the motivation for the work presented in this chapter. In particular, we propose to learn object class models for recognition from 3D computer aided design (CAD) models, as they are used in computer graphics applications, game design, or film making. These models can be obtained either from commercial providers, or downloaded free of charge from internet model sharing sites. In comparison to real world images, 3D CAD models have the advantage of being representable from arbitrary viewpoints, lighting conditions, and background scenes, providing an accurate description of object geometry, and explicitly separating shape from texture information. In addition, 3D CAD models are often highly structured, being composed of hierarchies of increasingly complex building blocks, sometimes including not only geometry information, but also physical constraints for animation.

The object class model presented in this chapter is an extension of the work presented in Chapter 5. It subsumes multiple instantiations of object class models of Chapter 5 in a single object class representation spanning multiple viewpoints. In comparison to Chapter 5, the individual object class models are improved by choosing robust local shape features in connection with powerful, discriminatively trained part detectors instead of spline features and representative part detectors.

7.1 INTRODUCTION

In the 70's and 80's the predominant approach to recognition was based on 3D representations of object classes (Nevatia and Binford, 1977; Marr and Nishihara, 1978; Brooks *et al.*, 1979; Pentland, 1986; Lowe, 1987). While being an intriguing paradigm these approaches showed only limited success when applied to real-world images. This was due both to the difficulty to robustly extract 2D image features as well as their inherent ambiguity when matching them to 3D models. Today, thirty years later, the predominant paradigm to recognition relies on robust features such as SIFT (Lowe, 2004) and powerful machine learning techniques. While enabling impressive results, e.g., for the PASCAL-VOC challenge (Everingham *et al.*, 2010), these methods have at least two inherent limitations. First, methods typically do not allow to recognize objects from arbitrary viewpoints but are limited to single viewpoints instead. And second, these approaches rely on the existence of representative and sufficient real-world image training data for object classes limiting their generality and scalability.

The starting-point of this chapter is therefore to go back to the idea of using 3D object models only and re-examine the problem of object class recognition from such 3D data alone, not using any natural training images of the object class. In contrast to early approaches, we draw from a multitude of advancements in both object class recognition and 3D modeling, which we use as tools for designing highly performant object class models. The first and most important tool is an abstract shape representation that establishes the link between 3D models and natural images, based on non-photorealistic rendering. The second tool is a collection of discriminatively trained part detectors, based on robust dense shape feature descriptors on top of this representation. The third tool is a powerful probabilistic model governing the spatial layout of object parts, capable of representing the full covariance matrix of all part locations. All three tools aim at capturing representative object class statistics from a collection of 3D models, increasing the robustness of the resulting object class models.

The main contributions of this chapter are as follows. First, we revisit the problem of object class recognition entirely based on 3D object models, avoiding the need for any natural training images of the objects. Second, we propose an abstract shape representation in connection with robust part detectors that establishes the link between 3D data and natural images. Third, we evaluate our model in a series of experiments with respect to multi-view detection and viewpoint classification (pose estimation), and demonstrate superior performance compared to state-of-the-art techniques on a standard multi-view recognition benchmark.

7.2 RELATED WORK

Recognition of 3D objects has a long history. While many of today's approaches model single 2D views rather than 3D objects, 3D object class models have been

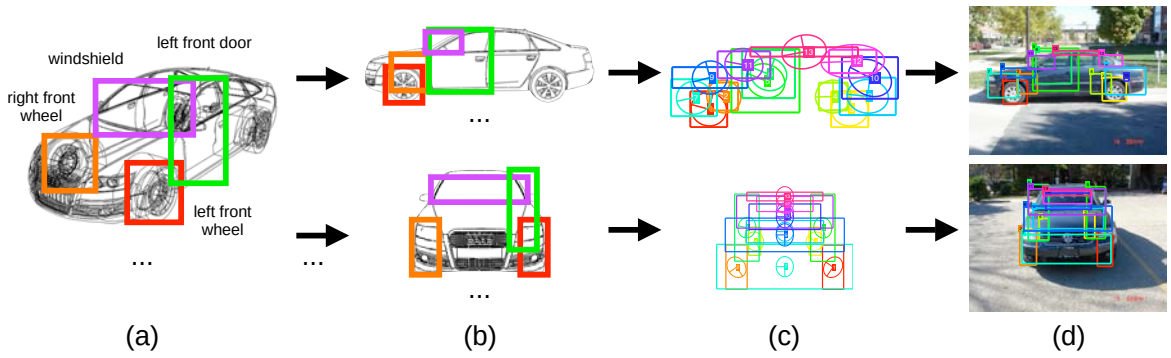


Figure 7.1: Learning shape models from 3D CAD data. (a) Collection of 3D CAD models composed of semantic parts, (b) viewpoint-dependent, non-photorealistic renderings, (c) learned spatial part layouts, (d) multi-view detection results.

revived recently as object recognition is inherently related to the object's three dimensional nature. 3D object class models are typically built either implicitly, by organizing training images according to their position on the viewing sphere, or explicitly, by establishing correspondences between training images and a given 3D geometry representative for an object class. In both cases, and in addition to representing 3D constraints, the robust encoding of object appearance learned from a sufficient amount of natural training images is considered key to success.

The first major line of research in 3D object recognition starts from a collection of natural training images depicting the object class of interest from varying viewpoints (Arie-Nachimson and Basri, 2009; Thomas *et al.*, 2006; Su *et al.*, 2009; Ozuysal *et al.*, 2009; Gill and Levine, 2009). The viewpoint itself is treated either as an observed (Gill and Levine, 2009; Liebelt and Schmid, 2010) or unobserved (Su *et al.*, 2009) variable, resulting in different amounts of supervision needed during training. Establishing correspondences between image features from different views by means of tracking (Thomas *et al.*, 2006) or imposing affine transformations (Su *et al.*, 2009) can then be used as the basis for rough estimates of three dimensional object geometry. These approaches have adopted sophisticated techniques to compensate for the large amount of required training data, such as sharing information between multiple codebooks by activation links (Thomas *et al.*, 2006), similarity transforms (Arie-Nachimson and Basri, 2009), or by synthesizing unseen viewpoints by means of a morphing variable (Su *et al.*, 2009). However, due to the reliance on sufficient training data from multiple viewpoints they are still bound to a typically coarse 3D and viewpoint representation of the object class, limiting the amount of variation captured by both appearance and geometry representations.

The second major line of research thus starts from a given 3D geometry representative for an object class, typically given in the form of one or a few 3D models (Liebelt and Schmid, 2010; Yan *et al.*, 2007), which is assumed to capture geometric variation better than a model built from a limited collection of viewpoint images. The geometry model then serves as a reference frame to which supplemental natural training image features are attached, which can then be matched to natural

images for recognition. While Yan *et al.* (2007) perform the attachment based on appearance similarity, Liebelt and Schmid (2010) establish the link between images and geometric model by spatial consistency. In particular, the geometric model is rendered from the same viewpoints as the training images (requiring viewpoint annotations). Both are overlaid the same regular grid, establishing correspondences between respective grid positions. However, these approaches still require a sufficient number of supplemental real-world training images again limiting their generality.

Rather than using real-world training images we go back to the idea of early papers (Nevatia and Binford, 1977; Marr and Nishihara, 1978; Brooks *et al.*, 1979; Pentland, 1986; Lowe, 1987) to use 3D object models alone. More specifically we resort to using 3D computer aided design (CAD) models exclusively, both for learning local shape and global geometry models for the object class of interest. Abandoning object training images altogether additionally circumvents the need for an attachment step, which is susceptible to introducing noise into the appearance representation. We note that the method by Liebelt *et al.* (2008) is also exclusively based on 3D CAD data, but has been superseded by Liebelt and Schmid (2010), which in turn is outperformed by our approach (see Sect. 7.5). Our work is different from Liebelt *et al.* (2008), in that we explicitly design an abstract shape representation for 3D CAD data that can be directly matched to natural images, while Liebelt *et al.* (2008) use photorealistic rendering techniques against varying backgrounds to produce features resembling the ones found in natural images. We further suggest a full covariance spatial model for capturing the geometric variation of a collection of 3D CAD models, while Liebelt *et al.* (2008) resort to a star model (via generalized Hough voting).

7.3 OBJECT CLASS REPRESENTATION

Our object class representation combines two prominent approaches. First, it represents object classes as an assembly of spatially arranged parts, which has been shown to be an effective strategy for dealing with intra-class variation and partial occlusion for generic object class recognition (Leibe *et al.*, 2006a; Fergus *et al.*, 2003). Second, it subsumes object classes in a collection of distinct models, where each model corresponds to a discrete viewpoint. For each viewpoint, the link between 3D CAD models used for training and natural test images is established by a local shape representation of object parts, based on non-photorealistic rendering.

7.3.1 Object classes as flexible part configurations

In the spirit of Fischler and Elschlager (1973), we choose a part-based object class representation as the basis for our approach. Instances of a given object class are assumed to consist of a fixed set of parts, subject to both constraints describing their spatial layout and their relative sizes. Following early uses of CAD models for recognition (Brooks *et al.*, 1979), but in contrast to recent work (Liebelt *et al.*, 2008;

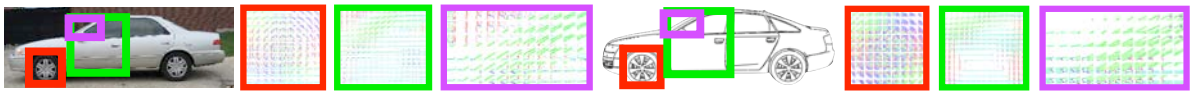


Figure 7.2: Comparison of shape representations fed into Shape Context descriptors for a real image (left) and a rendered 3D CAD model (right). For each colored bounding box, we show overlapping edge image patches, where edge orientation is encoded as hue and edge strength as saturation. Best viewed in color with magnification.

Liebelt and Schmid, 2010), we choose to not only use the three dimensional geometry from 3D CAD data, but additionally exploit included semantic information. In particular, we benefit from the fact that CAD data is typically created by human designers, often following intuitive routes when building complex models from simpler building blocks. As an example, consider the car model of Fig. 7.1 (a), which has been composed of semantically meaningful parts, such as wheels, doors, a roof, etc. While we cannot expect arbitrary 3D CAD models from the web to offer consistent part decomposition and labeling, we observe that all 41 car CAD models in our data base[11] share a common set of approximately 20 parts, from which we use 13 in our experiments (four wheels, both front doors, both bumpers, hood and trunk, windshield and rear window, the roof; see Fig. 7.1 (c)). Since both part decomposition and naming are potentially preserved in modern CAD file formats, we can establish semantic part-level correspondences between CAD models with minimal labeling effort. Inferring candidate parts and their correspondences automatically based on 3D geometry would be an alternative (Shalom *et al.*, 2008).

7.3.2 Viewpoint-dependent shape representation

In order to map 3D CAD data parts to the image plane, we apply a perspective projection according to the viewpoint of interest. In the image plane, each part is characterized by an axis-aligned bounding box (see Fig. 7.1 (a,b)). Note that we can still identify a part with a bounding box even in case it is not visible due to object-level self occlusion, as is the case for the right front wheel in the left side view of the car of Fig. 7.1 (b). In this case, the contents of the bounding box (orange) will depict the occluder (portion of the left front wheel, left front fender), not the originating object part. Following parts through occlusion in this fashion has the advantage of rendering occlusion reasoning superfluous, simplifying the design of the model. Coherence of part shapes between neighboring viewpoints also falls out naturally.

In contrast to earlier attempts at learning appearance models from 3D CAD data (Liebelt *et al.*, 2008), we choose a shape-based abstraction of object appearance at the core of our part-based representation. We focus on capturing edge information, which we expect to be repeatable across 3D CAD models of a given object class as

[11] Commercial models from www.doschdesign.com

well as natural images depicting instances of that class. At the same time, using the edge abstraction eliminates the need for rendering CAD models multiple times under varying lighting conditions, textures, and backgrounds, and having a learning algorithm finding out about relevant gradients afterwards. This intriguing property was shared by early approaches (Nevatia and Binford, 1977; Marr and Nishihara, 1978; Brooks *et al.*, 1979; Pentland, 1986; Lowe, 1987), but is often neglected by recent object class models. Specifically, we render three different types of edges for any 3D CAD model: crease edges, which are inherent properties of a 3D mesh, and thus invariant to the viewpoint, part boundaries, which mark the transition between object parts and often coincide with creases, and silhouette edges, which describe the viewpoint-dependent visible outline of an object (Hertzmann, 1999). In all three cases, we render edge strength (determined by the sharpness of the crease for crease edges) as well as orientation in the image plane.

In order to describe the contents of a part bounding box in the image plane, we use a specific flavor of Shape Context (Belongie *et al.*, 2000) descriptors that has proven to be highly robust in the context of object class detection in cluttered images (Andriluka *et al.*, 2009). These descriptors are densely sampled over a uniform grid of overlapping image patches, and accumulate edge orientations locally in log-polar histograms. Fig. 7.2 visualizes edge information fed into these descriptors for both a non-photorealistic rendering of a 3D CAD model (right) and a natural image (left). Please note the high degree of visual similarity between the two visualizations. It indicates that the chosen shape abstraction successfully captures common properties of both renderings and natural images, which we consider a key ingredient for robust recognition.

7.4 MULTI-VIEW OBJECT CLASS DETECTION

As outlined in Sect. 7.3, our multi-view object class detection framework is based on a set of distinct object class models, one for each particular viewpoint, which is sometimes referred to as a *bank of detectors* (Thomas *et al.*, 2006). All models are structurally equal, the only difference between them is the viewpoint-dependent data used for training. Final detection hypotheses are generated by combining hypotheses from the individual models.

7.4.1 Discriminative part shape detectors

In order to discriminate between object parts and image background, we use the highly performant part shape detectors proposed by Andriluka *et al.* (2009) in connection with the shape context features described in Sect. 7.3.2. For each object part, we train an Ada-Boost classifier (Freund and Schapire, 1997) on positive and negative training examples. Positive examples are obtained via non-photorealistic rendering of the object part in question. Negative examples are randomly sampled from a background image set, not containing the object class of interest. The set

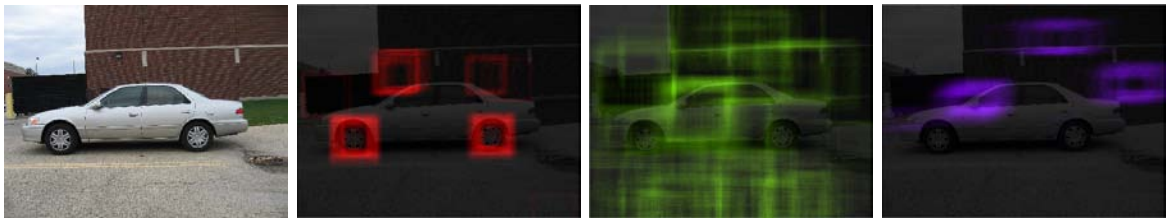


Figure 7.3: Part detector responses for *left front wheel* (red), *left front door* (green), and *windshield* (magenta), overlaid onto the original image (left). For each part, we show an accumulation of bounding boxes, weighted by detector response, drawn at the respective location and scale.

of positive training examples is further artificially enhanced by adding slightly translated and scaled (jittered) copies of the original examples. During detection, the trained classifier is evaluated in a sliding-window fashion at different image positions and scales. Fig. 7.3 gives example part responses for three different car parts in a left side view. We transform Ada-Boost classifier responses into pseudo-likelihoods using Platt scaling (Niculescu-Mizil and Caruana, 2005), and form a set of discrete candidate part locations (typically up to several 100K per part and image) by applying a threshold.

7.4.2 Probabilistic spatial model

Interestingly, most recent work in multi-view recognition has adopted star-shaped spatial models (Liebelt *et al.*, 2008; Su *et al.*, 2009; Arie-Nachimson and Basri, 2009; Yan *et al.*, 2007; Thomas *et al.*, 2006). In contrast to these prior works, our approach uses a more powerful probabilistic representation of the spatial layout of parts inspired by the constellation model (Fergus *et al.*, 2003). We use an efficient implementation along the lines of Chapter 5, from which we borrow the notation in the following paragraphs. The probabilistic formulation combines the shape of individual parts S, their relative scales R, and their overall spatial layout X. In contrast to Chapter 5, we do not model pair-wise (symmetry) relations between parts. During detection, the goal is to find an *assignment* of all P model parts to candidate part locations in a test image, denoted as the detection hypothesis $H = (h_1, \ldots, h_P)$. That is, h_p contains a candidate part identifier assigned to part p. The detection problem can be formulated as a maximum a posteriori (MAP) hypothesis search over the distribution $p(X, R, S, H|\theta)$, which is the joint posterior distribution of H and image evidence, given a learned model θ. It factors into separate likelihood contributions for local part shape, spatial part layout, relative part scales, and a (uniform) prior on hypotheses, as follows:

$$p(X, R, S, H|\theta) = \underbrace{p(S|H,\theta)}_{\textit{Local Shape}} \underbrace{p(X|H,\theta)}_{\textit{Layout}} \underbrace{p(R|H,\theta)}_{\textit{Relative Scale}} \underbrace{p(H|\theta)}_{\textit{Prior}} \tag{7.1}$$

Local part shape. In contrast to Chapter 5, where we model local shape $S(h_p)$ in a representative fashion as a spline curve governed by a Gaussian likelihood, we define $p(S(h_p)|\theta)$ to be a pseudo-likelihood from Platt-scaled Ada-Boost classifier responses, trained in a discriminative fashion (see Section 7.4.1). We again assume conditional independence between individual parts.

$$p(S|H,\theta) = \prod_{p=1}^{P} p(S(h_p)|\theta) \tag{7.2}$$

Spatial layout and relative scales. Spatial layout of parts is modeled in analogy to Chapter 5, as a joint Gaussian distribution over their coordinates $X(H)$ in a translation- and scale-invariant space (the *constellation*), using Procrustes analysis (Cootes, 2000). The model allocates independent Gaussians for the relative scale $R(h_p)$ of each part, i.e., the ratio between part and constellation scale.

$$p(X|H,\theta)\, p(R|H,\theta) = \mathcal{N}(X(H)|\theta) \prod_{p=1}^{P} \mathcal{N}(R(h_p)|\theta) \tag{7.3}$$

Learning and inference. Since we assume the densities for relative scales and spatial layout to be Gaussian, we can estimate parameters θ in a maximum likelihood fashion, given part-level correspondences. Following Chapter 5, we use an efficient Data-Driven Markov Chain Monte Carlo sampling algorithm for MAP inference. We approximate the MAP hypothesis $H_{\text{MAP}} = \arg\max_H p(H|X,R,S,\theta)$, which is is equivalent to $\arg\max_H p(X,R,S,H|\theta)$, by drawing samples from $p(X,R,S,H|\theta)$ using the Metropolis-Hastings (MH) algorithm (Gilks *et al.*, 1996). Employing the single component update variant of MH allows to separately update individual components of the target density, conditioned on the remaining portion of the current state of the Markov chain. This opens the possibility to guide the sampling towards high density regions by data-driven, bottom-up proposals (Zhu *et al.*, 2000; Tu *et al.*, 2005), which we instantiate by part shape likelihoods $p(S(h_p)|\theta)$.

7.4.3 Viewpoint estimation

In order to be able to detect potentially multiple object instances in an image, we run a number of independent Markov chains (typically 50) for each viewpoint-dependent detector of a bank. For each chain, we memorize the highest-scoring bounding box together with the viewpoint of the originating detector. We then apply a standard, greedy, overlap-based non-maximum suppression on all bounding box-viewpoint pairs, and retain all survivors as the final hypotheses concerning object bounding box and viewpoint.

7.5 EXPERIMENTAL EVALUATION

We evaluate the performance of our model on the *car* class of the 3D Object Classes data set introduced by Savarese and Fei-Fei (2007). The data set has been explicitly designed as a multi-view detection benchmark, containing 10 different cars, each pictured in front of varying backgrounds from 8 different 45 degree-spaced azimuth angles (*left*, *front-left*, *front*, *front-right*, *right*, *back-right*, *back*, *back-left*), 2 different elevation angles (*low*, *high*), and 3 different distances (*close*, *medium*, *far*). The resulting 48 viewpoints are typically not fully accurately met, but may be off by a few degrees in either direction. We evaluate object class detection from multiple viewpoints by first training an object class model consisting of a bank of 8 different detectors, where each detector corresponds to one of the approximate azimuth angles defined by the data set, using 41 3D CAD models. We expect our viewpoint-dependent detectors to be robust enough to cover both elevation angles. Similarly, varying distance is handled by considering part candidates at different scales. Fig. 7.5 visualizes 5 examples of the 8 learned models, together with corresponding example detections. It visualizes the part layouts of true positive detection hypotheses. In most cases, the hypothesized layout of object parts resembles the true layout pretty accurately, supporting exact localization at the object bounding box-level (by forming the smallest bounding box including all parts).

Comparison to state-of-the-art. We compare the performance of our model to three recent published results on the 3D Object Classes *Cars* data set, following the protocol of Su *et al.* (2009). Fig. 7.4 (a) gives precision/recall (P/R) plots for our bank of 8 detectors (green curve) and the methods of Su *et al.* (2009) (cyan curve), Gill and Levine (2009) (blue curve), and the very recent Liebelt and Schmid (2010) (magenta curve). Achieving an average precision (AP) of 81.0%, our method clearly outperforms all three related approaches (APs 55.3%, 72.6%, and 76.7%). Performance can be further improved by increasing the number of detectors to 36 in a 10-degree spacing (red curve, AP 89.9%, see dense viewpoint sampling for details).

Sensitivity to viewpoint variation. In a second experiment, we examine the sensitivity of our viewpoint-based object class model to discrepancies between viewpoints used for training and testing. For this purpose, we perturb the 8 original training viewpoints systematically by $p \in \{\pm 5, \pm 10, \pm 15, \text{and } \pm 20\}$ degrees, and test the performance of object class models consisting of all viewpoint-dependent detectors of a certain perturbation, amounting to banks of 16 detectors each (8 perturbed by $+p$ and 8 perturbed by $-p$ degrees). Fig. 7.4 (b) gives the corresponding P/R plots in different shades of red color, recapitulating the original green curve from Fig. 7.4 (a). We observe that, as expected, perturbation has a negative effect on performance in most cases, depending on the amount of perturbation. While for ± 10 degrees (light red curve), performance is on par with the original bank of 8 detectors (comparing the two curves; the AP of 81.2% is in fact even slightly higher), it drops significantly for ± 15 (dark orange curve, AP 70.3%) and ± 20 (light orange curve, AP 58.5%) degrees. Strikingly, even for ± 20 degrees, where all detectors are practically positioned as far away from the test image viewpoints as possible, the model still achieves an AP of 58.5%. The ± 5 detectors (dark red curve) improve (AP 81.3%) over the original bank of 8 detectors, managing to capture slight inaccuracies in the actual test image viewpoints.

Dense viewpoint sampling. In a third experiment, we want to determine the density of sampled viewpoints (VPs) required for good performance. We thus train banks of varying numbers of detectors, each bank representing a uniform sampling of the azimuth angle range of 360 degrees into equal size intervals. Fig. 7.4 (c) gives P/R curves for banks of detectors with interval sizes 5, 10, 15, 20, and 30 degrees (curves in shades of red and yellow color). We start sampling the azimuth angle range at 0 degrees (corresponding to a *left* side view) for each bank, and proceed counterclockwise from there. Note that this results in different numbers of sampled VPs coinciding with test image VPs for different banks. As a consequence, the evaluation involves both viewpoint density and number of coincident VPs. In Fig. 7.4 (c), we observe that an interval of 30 degrees (yellow curve, 4 coincident VPs) already provides a sufficient coverage of the azimuth angle range (AP 80.7%). Performance increases consistently for denser sampling and saturates at 10 degrees (light red curve, 4 coincident VPs, AP 89.9%, outperforming related work by 13.2%). An even denser sampling of 5 degrees does not further improve performance (dark

red curve, 8 coincident VPs, AP 89.8%).

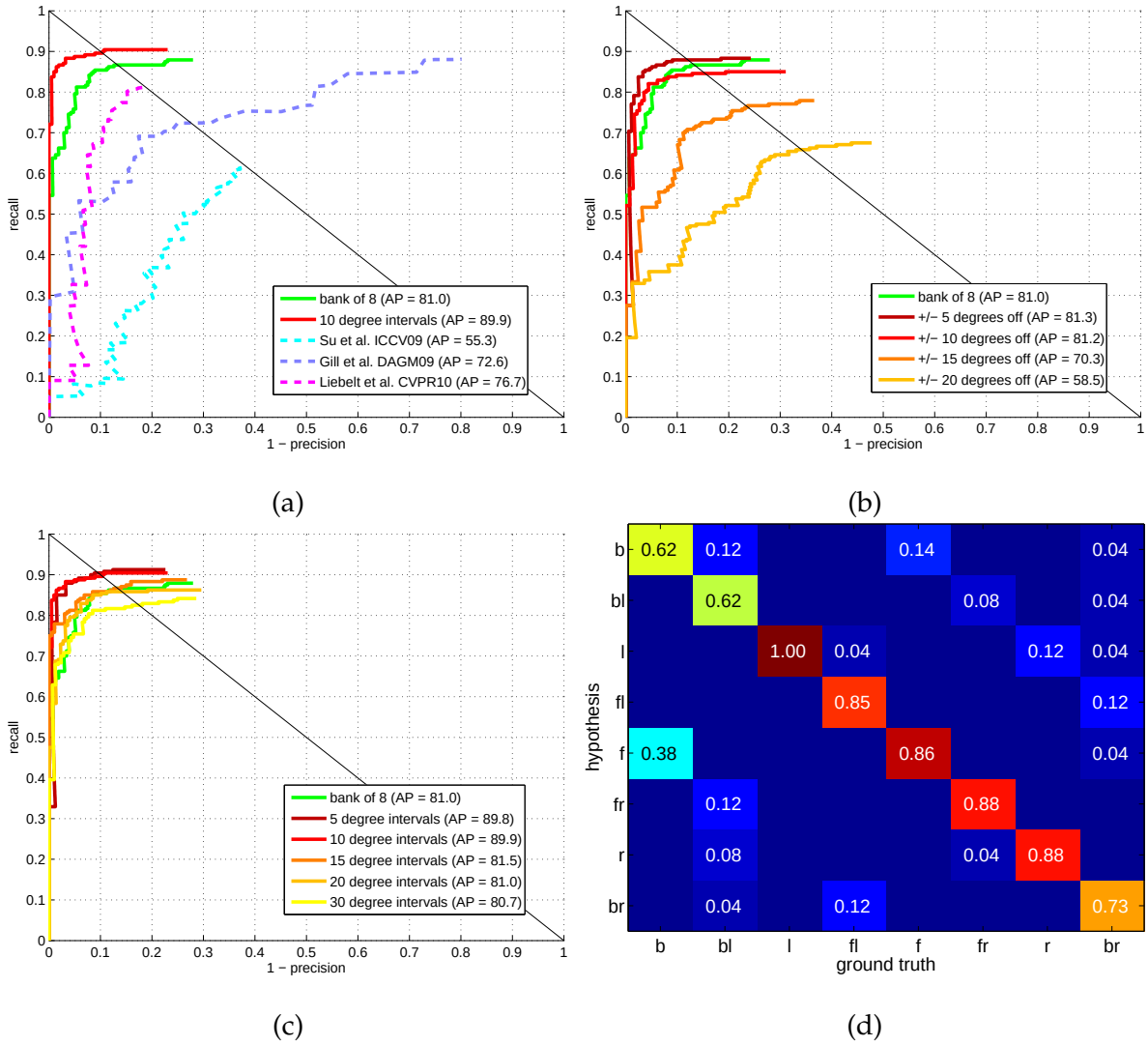


Figure 7.4: Multi-view object class detection results, (a) comparison to state-of-the-art (Su *et al.*, 2009; Gill and Levine, 2009; Liebelt and Schmid, 2010), (b) varying amounts of perturbation w.r.t. the true annotated viewpoint, (c) varying densities of sampled viewpoints, (d) confusion matrix for viewpoint classification.

Failure cases. We observe that missing recall is often caused by missing edge information due to low image contrast (dark car color, shadows), and occurs mostly for small scale objects pictured from the most distant (*far*) VP. This holds true for 91% of the cars missed by our best performing model.

Viewpoint estimation. Fig. 7.4 (d) gives the confusion matrix for classifying all true positive detections according to the 8 azimuth angles defined by the data set, using the bank of 8 detectors. While we observe that neighboring VPs are rarely

confused, confusion is larger for opposing views due to car symmetries (38% of *back* views are classified as *front* views). The average accuracy of 81% compares favorably to the best reported result of 70% by Liebelt and Schmid (2010).

7.6 CONCLUSIONS

In this chapter, we revisit the idea of learning shape models for object class recognition purely from 3D data, not using any natural training images of the object class of interest. While early approaches mostly failed in matching 3D models robustly to natural images, we benefit from intermediate advancements in object class recognition. By building our object class model on the robust combination of local part shape with a powerful model of spatial part layout, we demonstrate superior performance to state-of-the-art on a standard multi-view object class detection benchmark. While our current object class representation is based on individual per-viewpoint models, we expect integrating a continuous viewpoint estimate into a true unified 3D representation to be beneficial for performance.

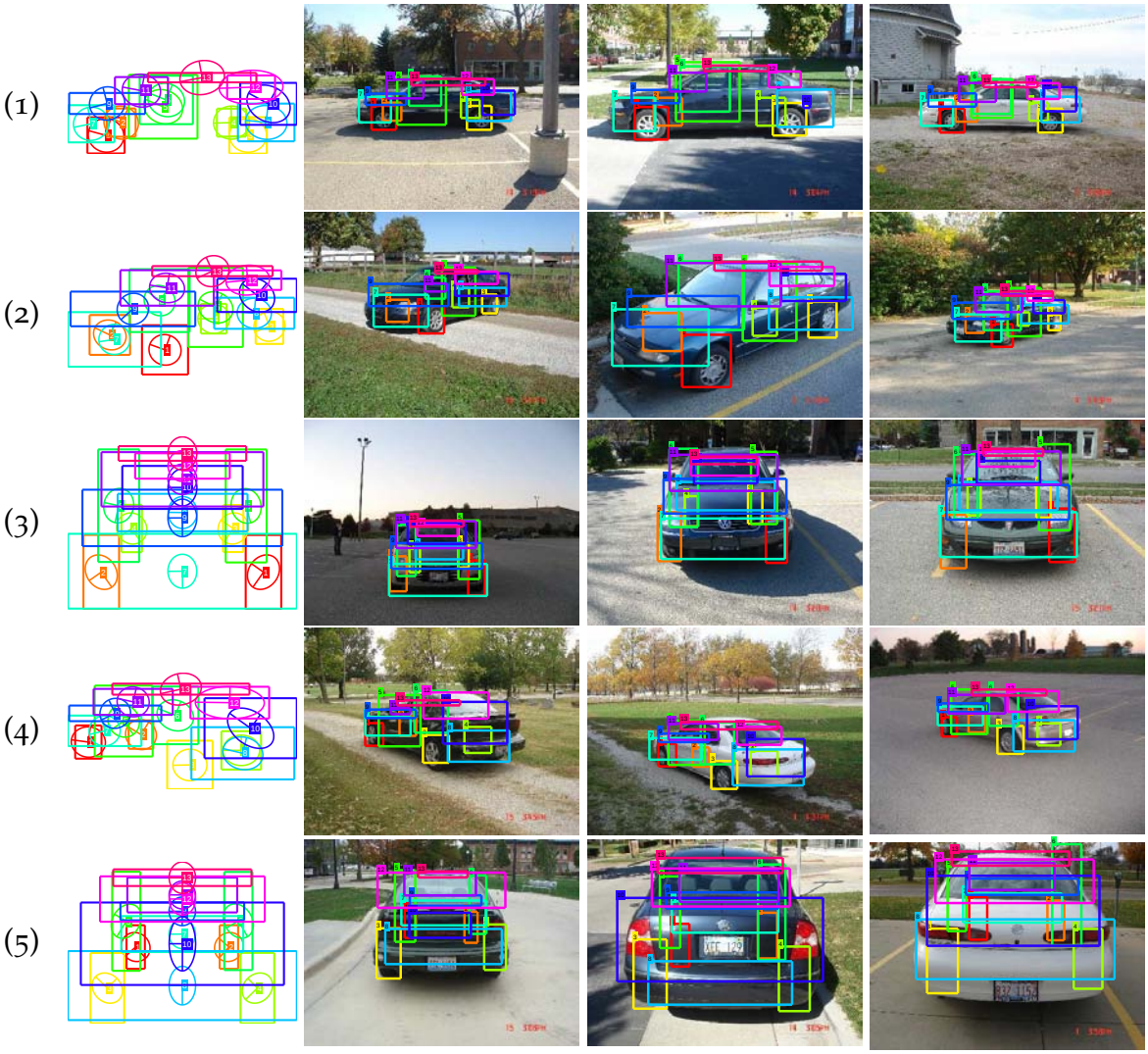


Figure 7.5: Viewpoint-dependent object class models for the viewpoints *left* (1), *front-left* (2), *front* (3), *back-left* (4), and *back* (5) (left-most column). Ellipses denote positional variance of parts, which are drawn at the learned mean scales. Example detections (right columns).

8 SEMANTIC RELATEDNESS FOR KNOWLEDGE TRANSFER

Contents

TRANSFERRING knowledge between object classes does not only require an appropriate representation of transferable knowledge, which has been the focus of Chapter 5, but also depends on the specification of potential sources and targets of transfer. While this specification has been given through manual supervision in Chapter 5, the present chapter concentrates on fully automatic methods. Since the focus is on automation rather than representational aspects of knowledge transfer, this chapter resorts to existing representations proposed by prior work, leaving the route towards purely shape-based methods adopted by earlier chapters.

As concerns the exploitation of additional sources of knowledge, this chapter draws from advances in natural language processing (NLP), by determining the semantic relatedness of object classes and likely visual properties by semantic relatedness measures, computed from linguistic knowledge bases, such as WordNet, Wikipedia, and different variants of general web and image search engines. The interface between linguistic and visual knowledge representations is realized by means of classifiers, relating named entities (object classes and visual properties) with visual features.

8.1 INTRODUCTION

Impressive recognition results were reported on a variety of object classes based on robust local features and powerful machine learning techniques. However, these approaches often rely on large amounts of training data limiting their scalability. It is clearly desirable to address the more challenging task of simultaneous recognition

of many object classes without the need for large training corpora. Reusing already acquired information by transferring knowledge between object classes has been suggested as a promising way to enable recognition of objects for which training data are scarce. The question what information to re-use in which context has been answered mostly by manual supervision (Lampert *et al.*, 2009) or by providing a few bootstrap training examples (Bart and Ullman, 2005b; Fink, 2004), limiting the applicability of these approaches.

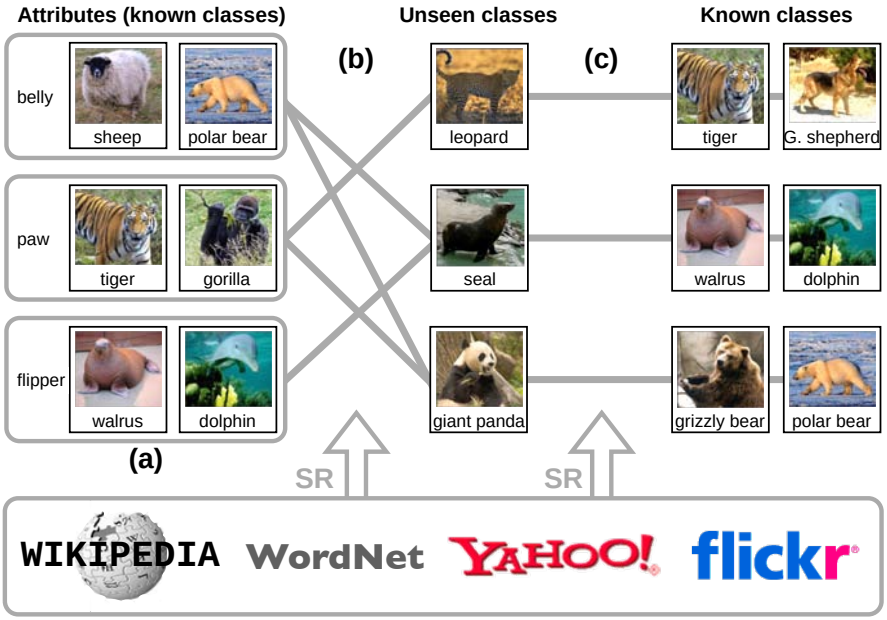


Figure 8.1: Inter-object class knowledge transfer. (a) Attribute inventory. Semantic Relatedness (SR) used to determine object class-attribute associations (b) and inter-object class similarity (c).

The main objective of our work is to extend such knowledge transfer approaches for object class recognition by significantly reducing the amount of needed manual supervision and training data. We do this by tapping into additional sources of information, namely linguistic knowledge bases, in order to provide the missing semantic link between sources (known object classes) and targets (unseen object classes) of knowledge transfer.

We choose two different models as the starting point of our work. In both models, knowledge transfer is realized by representing unseen object classes relative to known ones. The first model is based on an inventory of descriptive attributes (e.g. "belly", "paw", or "flipper", see Fig. 8.1(a)). For a given class, each attribute can be either active or inactive, resulting in a characteristic association signature for that class (e.g. "seal" is associated to "belly" and "flipper", see Fig. 8.1(b)). The second model is based on similarities between an unseen object class and known object classes (e.g. "leopard" is most similar to "tiger" and "G. shepherd", see Fig. 8.1(c)).

For both models we establish the semantic link between known and unseen classes by semantic relatedness (SR), which we measure using linguistic knowledge bases.

Based on these models, our study has two goals. The first goal is to better understand "how far we can get" in general with replacing manual supervision or seed training data by information acquired automatically from linguistic knowledge bases. The second goal is to evaluate the impact of particular choices of linguistic knowledge bases and semantic relatedness measures and to provide insights into their usefulness for different tasks. In contrast to most related work, we go beyond simple use of tags and image captions, and apply various state-of-the-art Natural Language Processing techniques which have, to our knowledge, not been used in a similar context in computer vision.

The main contributions of this chapter are as follows. First, we provide the missing semantic link for inter-object class knowledge transfer by using linguistic knowledge bases, based on two models (attribute-based and direct similarity-based). Second, for the attribute-based model, we explore different levels of automation in the knowledge transfer process. We not only determine the strengths of associations between object classes and attributes automatically using semantic relatedness, but also compile the attribute inventory automatically (see Fig. 8.1(a)). Third, we provide a rigorous experimental evaluation of different knowledge bases (such as WordNet (Fellbaum, 1998), Wikipedia, or the World Wide Web) and semantic relatedness measures and quantify their usefulness in the context of an object class recognition task. Fourth, we discuss the major differences, together with their possible reasons, between the examined knowledge bases and semantic relatedness measures. We believe that many of these insights are transferable to other vision tasks and may motivate a move-away from using WordNet as the default option for extracting semantic information.

The remainder of this chapter is organized as follows. After a review of related work (Sect. 8.2), we first introduce our model (Sect. 8.3) and the linguistic knowledge bases / semantic relatedness measures (Sect. 8.4) used in this study. We then give experimental results (Sect. 8.5) and conclude with an outlook (Sect. 8.6).

8.2 RELATED WORK

Due to the increasing need for scalable recognition, knowledge transfer between object classes has become an important topic in the vision literature. Amongst other directions, attribute-based representations have gained popularity recently by introducing an interpretable level of indirection between object classes (Ferrari and Zisserman, 2007; Kumar *et al.*, 2009; Wang and Forsyth, 2009). As attribute activations can characterize object classes without using reference exemplars they lend themselves to *zero-shot* classification of previously unseen object classes. Lampert *et al.* (2009) present zero-shot classification schemes based on attributes, where the associations between attributes and object classes are obtained using manual supervision by human subjects. Farhadi *et al.* (2009) advocate a paradigm shift from

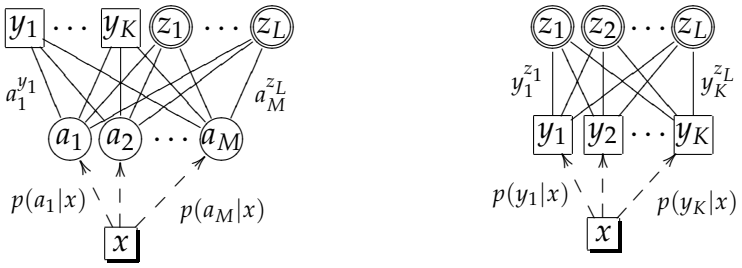
"naming" (by object classes) to "describing" (by attributes), distinguishing among common, discriminating, unusual, and unexpected attributes for object classes. In their work, zero-shot classification is phrased as a nearest-neighbor problem: a test image is classified as the most similar object class w.r.t. attribute descriptions. While the first of our object class representations is based on Lampert *et al.* (2009), we replace manual supervision with information extracted automatically from linguistic knowledge bases.

Other representations of transferable knowledge focus on more abstract notions of common or discriminating aspects between object classes, as practiced by the hierarchical classification schemes of Marszalek and Schmid (2007) and Zweig and Weinshall (2007). Similar in spirit, several works transfer knowledge in the form of learned distance metrics (Bart and Ullman, 2005a; Fink, 2004; Thrun, 1996) or object class priors (Fei-Fei *et al.*, 2006). Bart and Ullman (Bart and Ullman, 2005b) encode instances of previously unknown classes as collections of "familiar" classifier responses, i.e., similarities to known classes, and apply a nearest-neighbor scheme for classification. While the second of our object class representations is also based on such direct similarities, we extend this approach to zero-shot classification. That is we do not require the availability of any reference exemplar for unknown classes, but use information obtained from linguistic knowledge bases instead.

Obviously, using vision and language resources in combination promises mutual benefits for both and has been pursued actively in the literature. Approaches range from determining the "visualness" of language entities (Barnard and Yanai, 2006; Boiy *et al.*, 2008), over the construction of visually grounded ontologies (Popescu *et al.*, 2007; Wang *et al.*, 2008) to joint models of images, image tags, and image caption text (Barnard *et al.*, 2003; Li *et al.*, 2009). (Delezoide *et al.*, 2008) adds search engine hit counts for learning object-background co-occurrence statistics. While these approaches clearly demonstrate the benefit of using visual and language information together, most are still limited in the knowledge sources used (mostly WordNet or image tags) and the similarity measures applied (mostly path-length based measures). To our knowledge, our work is the first to give an in-depth exploration of both the possible knowledge bases and semantic relatedness measures for computer vision in general, and more specifically for knowledge transfer between object classes.

8.3 TWO MODELS FOR KNOWLEDGE TRANSFER

The main objective of our work is to automatically decide which knowledge to transfer between object classes by tapping into different language resources. We therefore extend two previous lines of work to enable zero-shot object class recognition, namely attribute-based recognition (Lampert *et al.*, 2009) and recognition based on direct similarities (Bart and Ullman, 2005b) between object classes. Both approaches model the relationship between known classes $y_1, \ldots, y_K$ and unseen classes $z_1, \ldots, z_L$. In attribute-based classification, an intermediate layer of descriptive attributes $a_1, \ldots, a_M$ serves as a level of indirection between known and unseen



(a) Attribute-based (Lampert *et al.*, 2009) (b) Direct similarity-based

Figure 8.2: Two models for zero-shot object classification. See Sect. 8.3 for discussions.

classes (see Fig. 8.2(a)). In recognition based on direct similarities, the known classes $y_1, \ldots, y_K$ serve directly as mediators for the unseen classes $z_1, \ldots, z_L$ (see Fig. 8.2(b)). The following subsections describes these models and our extensions to enable zero-shot learning in more detail.

8.3.1 Attribute-based classification

Attribute-based classification models object classes relative to an inventory of descriptive attributes. For a given class, each attribute can be either active or inactive, resulting in a characteristic association signature for that class. Fig. 8.2(a) gives a schematic overview of the Direct Attribute Prediction model (DAP) suggested by Lampert *et al.* (2009), which has been shown to yield better classification performance than Indirect Attribute Prediction (IAP).

Following the probabilistic formulation of the DAP model in Lampert *et al.* (2009), let $a^y = (a_1^y, \ldots, a_M^y)$ be a vector of binary associations $a_m^y \in \{0, 1\}$ between attributes a_m and training object classes y. A classifier for attribute a_m, trained by labeling all images of all classes for which $a_m^y = 1$ as positive and the rest as negative training examples, can provide an estimate of the posterior probability $p(a_m|x)$ of that attribute being present in image x. Mutual independence yields $p(a|x) = \prod_{m=1}^{M} p(a_m|x)$ for multiple attributes.

In order to transfer attribute knowledge to an unknown class z, we again assume a binary vector a^z for which $p(a|z) = 1$ for $a = a^z$ and 0 otherwise. The posterior probability of class z being present in image x is then obtained by marginalizing over all possible attribute associations a, using Bayes' rule $p(z|a^z) = \frac{p(a^z|z)p(z)}{p(a^z)} = \frac{p(z)}{p(a^z)}$:

$$p(z|x) = \sum_{a \in \{0,1\}^M} p(z|a)p(a|x) = \frac{p(z)}{p(a^z)} \prod_{m=1}^{M} p(a_m|x)^{a_m^z} \tag{8.1}$$

Assuming identical class priors $p(z)$ and a factorial distribution for $p(a) = \prod_{m=1}^{M} p(a_m)$,

we obtain

$$p(z|x) \propto \prod_{m=1}^{M} \left(\frac{p(a_m|x)}{p(a_m)} \right)^{a_m^z} \tag{8.2}$$

Attribute priors can be approximated by empirical means over the training classes $p(a_m) = \frac{1}{K} \sum_{k=1}^{K} a_m^{y_k}$ or set to $\frac{1}{2}$ (Lampert *et al.*, 2009). Classifying an image x according to test classes z_L uses MAP prediction $\text{argmax}_{l=1,\ldots,L} p(z_l|x)$.

This leaves us with estimating the class-attribute associations both for the known classes a_m^y as well as for the unknown classes a_m^z. In Lampert *et al.* (2009) human judgments of ten subjects (Kemp *et al.*, 2006; Osherson *et al.*, 1991) are used as the basis of these associations. While this has led to promising recognition results for unseen object classes, the main drawback of the approach is that it requires labor-intensive manual labeling to be applicable to a new domain (new sets of classes and attributes). As the main objective of our work is to reduce this dependency on human labeling effort, in the following subsections we discuss three extensions of the attribute-based model that alleviate this limitation: 1) mining the strengths of associations between classes and attributes by measuring their semantic relatedness using linguistic knowledge bases, 2) finding these attributes automatically, and 3) using the object classes themselves as *objectness attributes* (similarities of classes to each other), thereby eliminating the need for attribute-finding.

Mining object class-attribute associations. Our first extension of the attribute-based classification scheme taps into various language resources in order to automatically mine object class-attribute associations (see Fig. 8.1(b)). For this purpose, Sect. 8.4 introduces both different linguistic knowledge bases as well as several text-based semantic relatedness (SR) measures, which quantify the strength of relatedness between pairs of concepts (here class-attribute pairs). Note that we have to (manually) map full-text attribute descriptions designed for human comprehension (Kemp *et al.*, 2006; Osherson *et al.*, 1991) to concise terms which we can use as input to SR measures, which is inherently prone to drop information and is also susceptible to introducing noise (see Sect. 8.5).

Mining attributes. While mining class-attribute associations reduces the amount of manual supervision needed significantly, it is still relying on the definition of an appropriate inventory of descriptive attributes to start from. Ideally, these attributes should be able to discriminate between object classes (being associated to some but not all of them), provide sufficient coverage (all classes have at least a single attribute association), and be correlated to visual object class properties that can be observed in images. The creation of an appropriate attribute inventory is clearly a non-trivial task and has undoubtedly required careful engineering by Kemp *et al.* (2006); Osherson *et al.* (1991). Our second extension thus aims at avoiding this manual intervention by mining attributes from language resources (see Fig. 8.1(a)).

In our experiments, and in line with part-based modeling in the vision literature, we found *part attributes* (e.g. flipper for animals, wheel for vehicles) to meet the above described desired characteristics. Part attributes can be mined in various ways

from language resources, most obviously by using the explicit part relations encoded in WordNet. Here, we collect parts of all object class concepts of interest, including the parts of sub- and super-concepts recursively, resulting in 74 mined attributes[12] (compared to 85 manually defined ones).

Objectness as attributes. Our third extension uses object class names as attribute signifiers, characterizing the extent to which object classes are alike, which we casually denote *objectness*. "Giraffness", for example, then characterizes a group of object classes that are sufficiently similar to the giraffe class. As for attributes in general, several objectness attributes can be combined to yield a more precise description of an object class. Similarity is again determined from semantic relatedness measures (see Fig. 8.1(c)).

Interestingly, objectness attributes tend to encode complementary information to more generic ones, such as part attributes, by nature: while part attributes are often shared across diverse object classes, objectness attributes tend to form groups of classes that are highly related. As an example, consider the "giant panda" (bear) class in Fig. 8.1: it shares the "paw" attribute with diverse classes, such as "tiger" and "gorilla". On the other hand, similarity on the level of object classes yields "grizzly bear" and "polar bear" as the most similar classes.

8.3.2 Direct similarity-based classification

Similar in spirit to using objectness as attributes (see Sect. 12), but bypassing the level of indirection introduced by the attribute layer, we can use existing classifiers for known object classes $y_1, \ldots, y_K$ directly for zero-shot classification of unseen classes $z_1, \ldots, z_L$ (see Fig. 8.1(c), 8.2(b)). The particular choice of using a trained classifier for class y_k for classifying test class z_l depends on the similarity between the two, which can again be determined by using semantic relatedness measures.

This direct similarity model can be interpreted as a DAP model with $M = K$ attributes, where each attribute corresponds to exactly one training class y_k. We thus train classifiers for each class y_k to provide estimates of $p(y_k|x)$ for a test image x. In analogy to attribute-based classification (see Eq. (8.2)), the posterior of test image x is given as

$$p(z|x) \propto \prod_{k=1}^{K} \left(\frac{p(y_k|x)}{p(y_k)} \right)^{y_k^z}, \tag{8.3}$$

where y_k^z can be a binary association variable between the known class y_k and an unknown class z as for attribute-based classification (see Eq. (8.2)). We found empirically that using continuous weights is beneficial for performance and thus report results consistently for $y_k^z = w_{y_k}^z / \sum_{i=1}^{K} w_{y_i}^z$, assuming continuous weights w_y^z

[12] All software for computing object class-attribute associations from linguistic knowledge bases and obtained intermediate results (lists of mined attributes, object class-attribute associations) are publicly available on our web pages.

between z and y_k. Similarly, we restrict the considered classifiers to the $K = 5$ most similar ones in all experiments.

8.4 TEXT-BASED SEMANTIC RELATEDNESS

In this section, we describe the various linguistic knowledge bases and semantic relatedness (SR) measures that exploit natural language resources to gather information on the (visual) similarity of object classes or object class-attribute associations.[12] The most widely used resources in Natural Language Processing to calculate SR of concepts are without doubt WordNet (as the largest machine readable expert-created language ontology), Wikipedia (as the largest online encyclopedia), and the World Wide Web (as the largest public text collection one can use). For each of these resources we select a single, representative semantic relatedness measure designed to operate on that particular linguistic resource, given that the measure 1) has been widely used in previous studies, and 2) is generally regarded as competitive in terms of performance compared to other measures operating on the same resource.

WordNet and path length-based SR measures. Language ontologies and wordnets in particular are the most popular sources of machine readable information about a language, representing lexicalized concepts, synonymy, concept definitions, and various semantic relations. WordNet (Fellbaum, 1998) is a large scale lexical database of the English language, originally intended as model of human lexical memory. English words are organized into concepts (synonym sets or synsets) according to synonymy and various lexical and semantic relations are provided between these concepts. Due to its impressive size (over 100,000 concepts) and richness in encoded semantic relations, WordNet became the most important expert-created source of language information.

SR measures on WordNet (Budanitsky and Hirst, 2006) mostly use its graph structure (i.e., the encoded relations) to determine the path length between concepts or the shared information content of concepts. We use the similarity measure proposed by Lin (1998) that defines the similarity of two concepts c_1 and c_2 as $sim_{Lin}(c_1, c_2) = \frac{2*IC(lcs)}{IC(c_1)+IC(c_2)}$, where lcs denotes the lowest common subsumer of the two concepts in the WordNet hierarchy (i.e., the lowest common hypernym) and IC denotes the information content of a concept. IC is computed as $IC(c) = -\log p(c)$ where $p(c)$ is the probability of encountering an instance of c in a corpus. The probability $p(c)$ can be estimated from the relative corpus frequency of c and the probabilities of all concepts that c subsumes (Resnik, 1995).

Wikipedia and vector-based SR measures. Web based co-occurrence measures for SR often suffer from the noisy nature of web content. Wikipedia has been proposed as a source of background knowledge for calculating the semantic relatedness of words (Gabrilovich and Markovitch, 2007) and argued to provide a stable and noise-free resource w.r.t. this task.

Wikipedia is the largest online collaboratively built Encyclopedia, with more than 3 million articles for the English version. Wikipedia contains pages for concepts and each page provides a detailed and human edited description of the corresponding concept. In the past few years Wikipedia has been increasingly used as a source of world knowledge in Artificial Intelligence and Natural Language Processing in general and in text-based SR calculation in particular. Wikipedia-based SR measures are currently considered to be the state-of-the-art (Zesch and Gurevych, 2010).

The Explicit Semantic Analysis (ESA) measure of Gabrilovich and Markovitch used here represents each term as a vector of Wikipedia concepts (according to their use in the corresponding articles) and measures semantic similarity as the cosine of the corresponding concept vectors (thus capturing distributional similarity of the terms over Wikipedia): $sim_{ESA} = \frac{\overrightarrow{c_1} * \overrightarrow{c_2}}{\left|\overrightarrow{c_1}\right| * \left|\overrightarrow{c_2}\right|}$.

World Wide Web and hit-count based SR measures. Apparently the largest source of (textual) information is the World Wide Web itself. Because of this, search results of web search engines (i.e., search hit counts (HC) or text snippets) have been extensively used in many Natural Language Processing applications to model lexical semantic knowledge. With respect to SR, various measures have been proposed that use hit counts to measure term co-occurrence information as an indicator of term relatedness (Kilgarriff and Grefenstette, 2003). In our study, we used Yahoo to gather hit count information from the Web. From the many variants proposed in the literature we used the Dice coefficient to measure the similarity of two terms t_1 and t_2 as $sim_{DICE}(t_1, t_2) = \frac{HC(t_1,t_2)}{HC(t_1)+HC(t_2)}$, where HC represents the hit counts for a given term (or the term pair).

In connection with part attributes (see Sect. 8.3.1), we can refine web-based SR by making explicit use of part-whole (*holonym*) relations (Berland and Charniak, 1999). This is achieved by formulating web queries including *holonym patterns*, such as "elephant's tusks" or "patches of leopards". In particular, we use nine holonym patterns suggested by Berland and Charniak (1999) excluding "in" patterns, since these often denote non-visible object class properties or parts: (1-2) **whole***'s* **part**[*s*], (3-4) **whole***s'* **part**[*s*], (5-6) **part**[*s*] *of a* **whole**, (7-8) **part**[*s*] *of the* **whole**, (9) **part***s of* **whole***s*.

Web image search and hit-count based SR measures. The World Wide Web provides a natural opportunity to derive more visually oriented SR measures: using the same methods as described above, i.e., web search and Dice coefficient to calculate SR, we can restrict our search to image-related texts (captions, anchor texts, etc.) by using an image search engine like Yahoo Image Search or to human-assigned image tags and description using a collaborative photo management and sharing application like Flickr's search functionality. Performing image-related searches to approximate relatedness of concepts, we expect to get a more visually relevant relatedness measure.

Binarization and normalization. Since attribute-based zero-shot classification (see Sect. 8.3.1) requires binary associations between object classes and attributes, we binarize the continuous similarity values returned by SR measures as in Lampert *et al.* (2009), by a global threshold on the continuous-valued association matrix. The threshold is set to the mean of all matrix entries (we exclude the diagonal for objectness as it contains the similarity of a term with itself). This thresholding obviously requires the similarity values to be comparable across classes and attributes, which we achieve by normalization. We normalize matrix values by dividing by column and row sums prior to binarization.

8.5 EXPERIMENTS

In this section, we apply the various zero-shot classification schemes presented in Sect. 8.3 to a publicly available dataset, namely, the Animals with Attributes (AwA) dataset introduced by Lampert *et al.* (2009). In particular, we reproduce the previously reported results on attribute-based zero-shot classification relying on manual supervision, providing the basis for our evaluation. We then compare the performance of the different knowledge bases introduced in Sect. 8.4 in place of manual supervision and interpret the differences.

8.5.1 Experimental setup

The dataset consists of 50 mammal object classes, each containing at least 92 images, paired with a human provided attribute inventory and corresponding object class-attribute associations (Kemp *et al.*, 2006; Osherson *et al.*, 1991). We follow the experimental protocol of Lampert *et al.* (2009), using the provided split into 40 training and 10 test classes (24,295 training, 6,180 test images). We also use the provided pre-computed feature descriptors, namely, RGB color histograms, SIFT, rgSIFT, PHOG, SURF, and local self-similarity histograms. In contrast to Lampert *et al.* (2009) we concatenate all features to a single vector instead of training independent SVMs.

For computational reasons, we depart slightly from the protocol of Lampert *et al.* (2009) in our main experiments. First, we down-sample all training images to the minimum of 92 available images per class. Second, we use histogram intersection kernel SVMs (Chang and Lin, 2001) instead of χ^2 kernel SVMs. For better comparison to Lampert *et al.* (2009), we reproduce their results using all training images and χ^2 kernel SVMs as a reference and report the differences to our reduced setting.

8.5.2 Experimental results

Tab. 8.1 gives zero-shot classification results in the form of area under ROC curve (AUC) scores for the ten individual test classes (first ten columns) and their average (last but one column). The last column gives the corresponding average multi-class

Source (Measure)	Area under ROC curve (AUC) in %										mean AUC (in %)	mean accuracy (in %)
	chimpanzee	giant panda	leopard	persian cat	pig	hippopatamus	humpback whale	raccoon	rat	seal		
1. Reproduction of the results in Lampert et al. (2009))												
all images, χ^2	86	65	88	84	73	77	99	78	76	78	80.3	40.3
92 images, hist. int.	87	64	86	84	71	75	98	72	71	77	78.5	34.7
2. Mined object class-attribute associations												
WordNet (Path)	53	52	50	**79**	60	48	86	57	70	51	60.5	15.5
Wikipedia (Vector)	58	65	74	78	62	**70**	88	**73**	63	65	69.7	**27.0**
Yahoo Web (HC)	39	65	63	49	**73**	52	91	39	**76**	56	60.4	22.2
Yahoo Img (HC)	74	**77**	**81**	61	72	57	**97**	63	53	**76**	**71.0**	26.3
Flickr Img (HC)	**79**	72	**81**	76	68	56	94	64	47	63	70.1	20.0
Yahoo Img & Flickr	78	77	83	69	72	57	97	64	49	70	71.6	27.8
3. Mined attributes (and associations)												
WordNet (Path)	49	62	59	**70**	55	43	86	63	58	52	59.8	17.8
Wikipedia (Vector)	65	61	**69**	68	61	70	93	59	58	56	66.0	19.7
Yahoo Web (HC)	39	71	45	47	70	53	95	39	**65**	49	57.4	19.5
Yahoo Img (HC)	66	**79**	66	63	47	58	**98**	59	62	54	65.2	**23.6**
Flickr Img (HC)	60	70	67	60	49	63	**98**	**67**	60	51	64.6	22.9
Yahoo Holonyms (HC)	**78**	61	68	59	**71**	**77**	**98**	66	60	**59**	**69.9**	21.5
Wikipedia & Yahoo Hol.	77	65	74	68	70	76	98	65	63	61	71.6	26.9
4. Objectness as attributes												
WordNet (Path)	**79**	**67**	71	72	**70**	70	97	61	63	62	71.2	25.6
Wikipedia (Vector)	67	48	65	69	52	67	94	68	62	72	66.4	23.3
Yahoo Web (HC)	51	65	69	83	66	**76**	98	52	57	51	66.7	25.5
Yahoo Img (HC)	68	63	76	**86**	67	65	**99**	**77**	**71**	71	**74.1**	**33.0**
Flickr Img (HC)	57	59	**78**	–	63	66	98	70	**71**	**74**	*68.5*	*11.2*
WordNet & Yahoo Img	81	70	80	80	73	70	98	68	71	70	76.1	33.5
5. Direct similarity												
WordNet (Path)	**88**	73	82	59	60	68	**98**	67	66	73	73.4	29.7
Wikipedia (Vector)	79	**77**	84	82	68	60	**98**	74	77	67	76.6	33.2
Yahoo Web (HC)	84	72	**88**	82	**77**	**70**	**98**	**76**	71	60	77.7	34.7
Yahoo Img (HC)	85	70	78	**85**	**77**	64	**98**	73	77	**81**	**78.8**	**35.7**
Flickr Img (HC)	84	72	86	78	**77**	63	**98**	72	**78**	73	77.8	32.5
Yahoo Img & Flickr	84	71	82	84	77	63	98	73	78	78	78.9	33.9
all	84	75	84	84	77	65	98	74	78	79	79.7	34.1

Table 8.1: Zero-shot classification results on the AwA data set (Lampert *et al.*, 2009). The best results per table section are given in bold font. "–" denotes unavailable results due to all-inactive attributes, set to chance level = 50% for mean calculation. See Sect. 8.5 for discussions.

classification accuracies. Each row of the table corresponds to a single classification experiment. The row-wise sections of the table mark the different variants of knowledge transfer (Sect. 1: reproduction of the results of Lampert *et al.* (2009), Sect. 2: attribute-based classification using mined object class-attribute associations, Sect. 3: attribute-based classification using mined attributes, Sect. 4: attribute-based classification using objectness as attributes, Sect. 5: direct similarity-based classification). Each row-wise, numbered section (except Sect. 1) gives results for various knowledge bases in the same consistent ordering.

1. Reproduction of the results in Lampert *et al.* (2009). Our implementation, closely following the settings in Lampert *et al.* (2009), using all available training images and χ^2 kernel SVM, achieves an average AUC of 80.3% and corresponding multi-class classification accuracy of 40.3% (Tab. 8.1, Sect. 1). This is very close to 80.7% and 40.5% reported in Lampert *et al.* (2009), respectively. Down-sampling the training set to 92 images per class and using histogram intersection kernel SVM decreases performance slightly, but not significantly, to 78.5% and 34.7%, respectively (Tab. 8.1, Sect. 1). All results that follow are based on this computationally more manageable setting.

2. Mined object class-attribute associations. We start by comparing the performance of the various knowledge bases using the original set of attributes proposed by Lampert *et al.* (2009), using semantic relatedness to determine class-attribute associations (Tab. 8.1, Sect. 2). We observe that Yahoo Img performs best on average (71.0% AUC), closely followed by Flickr Img (70.1%). This is expected since both are based on image-related texts and inherently capture important correlations between terms and visual attributes, which can be beneficial for recognition. The difference between Yahoo Img and Flickr Img is minor and may be in consequence of the generally smaller coverage of Flickr compared to the full web used by Yahoo Img. Similar performance is achieved by Wikipedia (69.7%). We attribute this to Wikipedia's encyclopedic nature which provides concise and noise-free explanations of concepts.

Last are WordNet (60.5%) and Yahoo Web (60.4%). They show a significant drop in performance ($\approx$ 10%) compared to the first three knowledge bases. For Yahoo Web this drop is due to an increased level of noise compared to image search or Wikipedia. In particular, we observed incidental co-occurrences on web pages and polysemous expressions (e.g. attribute term "pad", class term "seal") to have a negative effect on performance. Although incidental co-occurrences do not exist in WordNet, it does suffer from (non-disambiguated) polysemous terms. However, more important is the fact that path lengths are a poor indicator of semantic relatedness between object class and attribute concepts, as they are computed from hypernym relations: since object classes and attributes are inherently different in nature, they are likely to lie in entirely different subtrees of the hypernym hierarchy. Consequently, path length is no longer representative for their semantic relatedness.

In an attempt to benefit from potentially complementary information from two

different knowledge bases, we report additional results for the fusion of the two best competitors, Yahoo Img and Flickr Img. Fusion is performed by multiplying the respective class probabilities from both knowledge bases. In fact, this fusion performs slightly better than either individual knowledge base (71.6%).

In general, using knowledge bases for acquiring object class-attribute associations (at best 71.6%) performs consistently worse ($\approx 7\%$) than using manually defined associations from Lampert *et al.* (2009) (78.5%). We acknowledge that this performance drop is significant, but stress that the information provided to the human judges is far more descriptive than the simplified terms used to query the knowledge bases. E.g. "nest" abbreviates the attribute "keeping their young in a designated, enclosed area". We consider the obtained results, in connection with the significantly reduced amount of manual supervision, highly encouraging and an important contribution towards scalable recognition.

Above results are also reflected in the mined class-attribute associations. Although it is not always easy to judge if the associations are meaningful, we provide an example for the visual attribute "striped" to give an impression of the quality of mined similarities. In the following we list the four top ranked mammal classes for "striped" in decreasing order: Manual: *zebra, tiger, skunk, raccoon*; WordNet: *elephant, seal, mouse, bat*; Wikipedia: *zebra, skunk, tiger, Chihuahua*; Yahoo Web: *zebra, collie, Dalmatian, polar bear*; Yahoo Img: *zebra, skunk, tiger, Persian cat*; Flickr Img: *skunk, tiger, zebra, leopard*.

3. Mined attributes (and associations). Sect. 3 of Tab. 8.1 gives results for using automatically mined (part) attributes instead of manually defined ones, as presented in Sect. 2 of Tab. 8.1, fully avoiding any kind of manual supervision. Disregarding Yahoo Holonyms, we found Wikipedia, Yahoo Img, and Flickr Img again to perform best (66.0%, 65.2%, 64.6%) with a significant margin to the next best knowledge bases (WordNet 59.8%, Yahoo Web 57.4%). This is consistent with the results for manually defined attributes and underpins the differences between the knowledge bases highlighted above.

Moving from manually defined to automatically mined attributes results in a general drop in performance for the measures discussed above. We attribute this to the reduced number of attributes (85 to 74) and the reduced diversity of attributes (colors, context, parts etc. versus parts only).

For automatically mined attributes we give additional results for a specific flavor of Yahoo Web which allows to formulate queries specifically tailored towards part attributes (Yahoo Holonyms, see end of Sect. 8.4). The resulting semantic relatedness estimates clearly benefit from this specificity, obtaining the best performance of 69.9%, which is comparable to the previously reported results for manually defined attributes. The fusion of the best two, Wikipedia and Yahoo Holonyms, is again beneficial (71.6%) and on par with manually defined attributes.

4. Objectness as attributes. Sect. 4 of Tab. 8.1 gives results for using objectness as attributes for attribute-based classification. Specifically, we use all 50 class names as

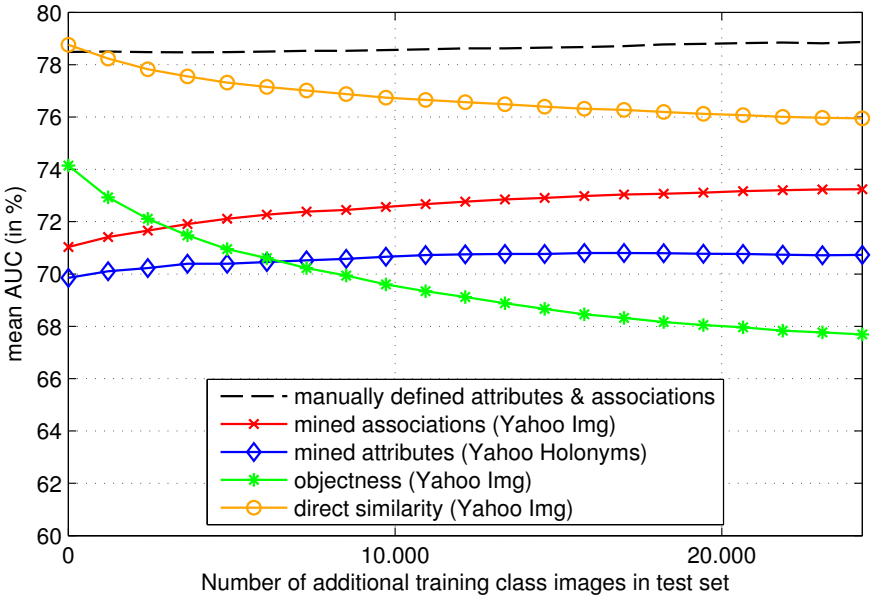


Figure 8.3: Adding training class images to the test set.

attribute signifiers. Yahoo Img is again best (74.1%). In contrast to sections 2 and 3, WordNet (71.2%) gains relative performance and performs second best: using objects (objectness) as attributes renders the corresponding concepts perfectly comparable to object classes with respect to the hypernym hierarchy. Path length becomes a valid measure of semantic relatedness. Next are Yahoo Web (66.7%) and Wikipedia (66.4%). Although Flickr Img performs on average slightly better (68.5%), it could not yield results for all test classes due to a lack in coverage: the user provided text content does often not provide sufficient statistics for co-occurring object classes. The fusion of the best two, WordNet and Yahoo Img is again beneficial (76.1%).

5. Direct similarity. Sect. 5 of Tab. 8.1 gives results for zero-shot classification based on direct similarity between object classes. We observe that most knowledge bases show very similar and sometimes even slightly better performance compared to using manually defined object class-attribute associations (Yahoo Img 78.8%, Flickr Img 77.8%, Yahoo Web 77.7%, Wikipedia 76.6%, fusion Yahoo Img & Flickr 78.9%, fusion of all 79.7%). Only WordNet performs worse (73.4%). This is in line with the observation that for a given test class, the 5 chosen most similar classifiers used for zero-shot classification are quite reliable on average and generally similar among the different knowledge bases. Even more importantly, the direct similarity model circumvents the need for an intermediate attribute layer, effectively eliminating one potential source of noise from the process: while a specific classifier used in direct similarity-based classification is guaranteed to be trained from appropriate training data, the training set of an attribute classifier is determined using (noisy) semantic relatedness measures.

Training class images in the test set. Although the data set and testing protocol suggested by Lampert *et al.* (2009) provides a valuable resource for zero-shot classification experiments, it exhibits a distinct property that may lead to a biased view on the results. Namely, the set of object classes used for training and test is disjoint: a zero-shot classifier under test is never challenged to distinguish a previously unseen class from one it actually knows. We expect this distinction to be difficult, since it asks for classifying as negatives those classes that have been used as positive examples during training.

We visualize this effect in Fig. 8.3, where we plot AUC for the best performing methods per section of Tab. 8.1, varying the number of training class images which we inject as negatives to the test set. The horizontal axis denotes the number of injected training class images not actually used for training (these images exist due to the down-sampled training set). As expected, we observe performance drops for objectness (Yahoo Img) and direct similarity (Yahoo Img) for added negatives. In contrast to this, all classifiers based on generic attributes (mined associations (Yahoo Img), mined Attributes (Yahoo Holonyms), and manually defined attributes & associations) seemingly generalize better over added negatives, which we explain by their complementary nature (see Sect. 8.3.1). They tend to form groups of more diverse object classes than objectness attributes, and thus limit the influence of specific (positive) training classes appearing as negatives in the test set.

Summary. In summary, we conclude that Yahoo Img is unparalleled with respect to coverage and precision. It outperforms most other knowledge bases in attribute-, objectness-, and direct similarity-based classification approaches, reaching a level of performance comparable to manually defined object class-attribute associations for direct similarity-based classification. Flickr Img is always inferior to Yahoo Img due to smaller coverage. Wikipedia achieves similar performance as web image search. Yahoo Web and WordNet provide competitive results only for inter-object similarity. Fusion of complementary knowledge bases always helps. Considering the different zero-shot classification approaches, objectness as attributes and direct similarity-based classification are superior to both manually defined and automatically mined attributes and can even level manually provided object class-attribute associations.

8.6 CONCLUSIONS

Having recognized knowledge transfer as a promising route to scalable recognition, we believe that further reducing the needed amount of manual supervision is a vital ingredient for making it a widely applicable tool for computer vision. In this chapter, we demonstrate that manual supervision can in principle be fully replaced by tapping into linguistic knowledge bases, by the example of zero-shot object class recognition. Although this leads to a decrease in classification accuracy for attribute-based classification, direct similarity-based classifiers achieve performance on par with manual supervision. In our evaluation, we observe that the different characteristics

of knowledge bases often result in largely different results. In particular, Yahoo image search and Wikipedia are always among the best choices while Yahoo web search and WordNet are especially inferior for attribute-based associations.

9 CONCLUSIONS AND FUTURE PERSPECTIVES

Contents

Motivated by the good performance of modern object class recognition systems on the level of individual object classes, the simultaneous recognition of many classes is coming into view (e.g., in the ImageNet Large Scale Visual Recognition Challenge [13] (Everingham *et al.*, 2010)). Scaling recognition to higher numbers of classes is difficult, however, due to both increasing model complexity and the need for increasing amounts of training data. Knowledge transfer between object classes has been identified as a promising route to enable scalable recognition (Torralba *et al.*, 2004; Fink, 2004; Bart and Ullman, 2005b; Fei-Fei *et al.*, 2006), by re-using once acquired information in the context of related recognition tasks. In particular, learned models for object classes (Fink, 2004; Bart and Ullman, 2005b) or more generic visual properties (Lampert *et al.*, 2009; Farhadi *et al.*, 2009) can be recombined to facilitate the learning of object classes with scarce training data. Besides increasing scalability, knowledge transfer enables novel tasks, such as recognizing object classes for which no training data are available (zero-shot recognition (Lampert *et al.*, 2009)). In this case, missing training data is compensated by using additional sources of knowledge. Training data is replaced by an abstract description that characterizes the object class of interest relative to existing model building blocks. Based on these encouraging prospects, this thesis has explored four different dimensions of knowledge transfer in object class recognition.

First, we investigated the role of visual features as a low level representation of transferable knowledge. Based on an extensive evaluation of existing state-of-the-art local feature detectors and descriptors, we identified shape-based features in connection with powerful spatial models as a promising candidate representation. Building upon this result, we further introduced a novel flavor of local shape-based features, as well as a generic appearance descriptor based on shading artifacts.

Second, we highlighted the connection between knowledge transfer and gen-

[13] http://www.image-net.org/challenges/LSVRC/2010/

eralization across basic-level object categories (Rosch *et al.*, 1976), by recognizing objects according to potential functions or affordances (Gibson, 1977). In particular, we demonstrated that visually distinct hints on affordances, modeled as collections of local shape features, can be shared and hence transfered between object classes.

Third, we designed shape-based object class models for knowledge transfer, representing object classes as spatially constrained assemblies of parts, including pair-wise symmetry relations. These models are both compositional and incremental, allowing for knowledge transfer either on the level of entire object class models or restricted to a subset of model components. While knowledge transfer in these models has to be guided by manual supervision, we demonstrated the benefit of knowledge transfer for object class recognition when learning from scarce training data.

And fourth, we demonstrated that exploiting additional sources of knowledge besides real world training images can aid object class recognition, effectively transferring knowledge between different representations. In particular, we used linguistic knowledge bases in connection with semantic relatedness measures to automatically determine potential sources and targets of knowledge transfer for zero-shot recognition, and showed the successful learning of shape-based object class models from collections of 3D computer aided design (CAD) models, not using any real world training images of the object class of interest.

In summary, this thesis has achieved encouraging results with respect to four different dimensions of knowledge transfer, namely, specialized visual feature representations, generalization across basic-level categories, compositional object class models, and the exploitation of additional sources of knowledge, confirming the benefits of knowledge transfer. We consider the presented work along these individual dimensions valuable contributions to the field of knowledge transfer in object class recognition, advancing the current state-of-the-art with respect to both representing transferable knowledge and determining potential sources and targets of knowledge transfer automatically. As a side effect, we were able to obtain object class recognition results often superior to or en par with prior work. While each of the presented contributions leaves room for improvement (see Section 9.2), the biggest potential for future work lies in their combination. We believe that knowledge transfer can be an important ingredient for the future development of scalable and powerful recognition methods, in connection with rich representations, the exploitation of a multitude of knowledge sources, and clever strategies for the explorative analysis of visual scenes.

9.1 DISCUSSION OF CONTRIBUTIONS

In this thesis, we have explored knowledge transfer in object class recognition along various dimensions. The resulting contributions span both general object class recognition and contributions specific to knowledge transfer.

9.1.1 Contributions to object class recognition in general

As concerns general object class recognition, we have presented an extensive evaluation of existing state-of-the-art local feature detectors and descriptors, targeted at the specific task of object class recognition (Chapter 3). The focus has been on shape-based features and object classes that are characterized by shape rather than appearance, and comprised the analysis of the corresponding feature spaces in isolation as well as in combination with different image classification frameworks and varying amounts of spatial information. As a result, we concluded that low-dimensional shape-based features, and k-AS (Ferrari *et al.*, 2008) in particular, can outperform other features for added spatial information, despite their weak discriminative power when compared in isolation. The evaluation was conducted on both the publicly available standard recognition benchmark Caltech-101 and two newly proposed data sets of shape-based object classes, which we have made publicly available, and which since have been used by various researchers (Fidler *et al.*, 2009; Torki and Elgammal, 2010).

In Chapter 5, we have designed a probabilistic, part-based object class model for knowledge transfer, extending the constellation model of Weber *et al.* (2000); Fergus *et al.* (2003). Motivated by the success of local shape features in the evaluation of Chapter 3, the model was built upon a novel flavor of local shape features inspired by Ferrari *et al.* (2008), and added symmetry relations between pairs of parts (Brady and Asada, 1984) to the original constellation model formulation. While adding symmetry relations effectively increased the representational capacity of the model, we could scale up the number of image features the implementation could handle from dozens reported by Fergus *et al.* (2003) to several thousands, by adopting a data-driven Markov Chain Monte Carlo (DDMCMC) sampling scheme for approximate MAP inference. As a result, we demonstrated superior performance of our model compared to two prior methods (Ferrari *et al.*, 2007; Fritz and Schiele, 2008) on a standard shape-based recognition benchmark (ETHZ Shape Classes (Ferrari *et al.*, 2006b)).

We further extended the purely shape-based object class model of Chapter 5 by adding a novel type of semi-local appearance descriptor, modeling shading artifacts in real world images (Chapter 6). More specifically, we introduced a physically inspired model of the observed shading on cylindrical surface primitives (Weinshall, 1992; Haddon and Forsyth, 1998), phrased as an energy minimization problem. In our experiments on the ETHZ Shape Classes data set, we confirmed the validity of our formulation on real world images, its robustness to specular effects on non-Lambertian surfaces, and its potential to improve object class recognition performance.

While the object class model of Chapter 5 was limited to single-view recognition, we proposed an extension to multiple viewpoints in Chapter 7. Following the bank of detectors paradigm (Thomas *et al.*, 2006), the corresponding object class representation subsumed an object class in a collection of distinct, structurally equal, viewpoint dependent models. Each of these models combined discriminatively

trained part-shape detectors (Belongie *et al.*, 2000; Andriluka *et al.*, 2009) with the powerful probabilistic spatial model of Chapter 5. In contrast to prior work, we demonstrated that highly performant multi-view object class models can be learned entirely from synthetic data, namely, from 3D CAD models, without using any real world training images of the object class of interest. We attributed the performance of our approach to a novel representation of local object shape, based on non-photorealistic rendering techniques. In our experiments on the publicly available 3D object classes data set Savarese and Fei-Fei (2007), we showed superior performance of our model compared to prior work (Gill and Levine, 2009; Su *et al.*, 2009; Liebelt and Schmid, 2010).

9.1.2 Contributions specific to knowledge transfer

As concerns knowledge transfer, we have shown its connection to generalization across basic-level categories in Chapter 4. More specifically, we have approached functional object class recognition based on object affordances from a different angle than prior work (Stark and Bowyer, 1991; Stark *et al.*, 1993; Green *et al.*, 1995; Rivlin *et al.*, 1995; Bogoni and Bajcsy, 1995), by starting from an existing object class recognition framework (the implicit shape model (Leibe *et al.*, 2006a)), in connection with robust local shape features (Ferrari *et al.*, 2008) and learning by observation. As a result, we demonstrated the transferability of affordance cues, visually distinct hints on object affordances, across objects belonging to different basic-level categories, on real world images of the ETHZ Shape Classes benchmark (Ferrari *et al.*, 2006b).

In Chapter 5, we proposed a probabilistic, part-based object class model for the specific purpose of knowledge transfer. The model was designed to be both compositional, allowing for the explicit transferability of constituent model parts, and incremental, supporting to incorporate prior information in the spirit of Bayesian priors (Fei-Fei *et al.*, 2006). In our experiments, we demonstrated the benefit of using transfered knowledge in addition to scarce training data on a newly proposed data set of animal object classes, for both the transfer of entire object class models and the transfer of proper subsets of parts.

In a first attempt to exploit additional sources of knowledge for object class recognition, we demonstrated the learning of multi-view object class models from 3D CAD data in Chapter 7. At the core of our approach, we introduced a novel, shape-based representation of object parts, which allowed to interface between 3D CAD models and real world images. This representation was based on edge information, which we extracted from 3D CAD data (part boundaries, crease edges, silhouette edges) and real world images (Canny edges), in connection with a robust encoding in the spirit of shape context (Belongie *et al.*, 2000). In combination with discriminatively trained part detectors (Freund and Schapire, 1997; Andriluka *et al.*, 2009), our experiments showed the superiority of our representation compared to approaches based on real world images alone (Gill and Levine, 2009; Su *et al.*, 2009) and in combination with 3D models (Liebelt and Schmid, 2010).

Chapter 8 constitutes our second attempt to exploit additional sources of know-

ledge, this time in the form of linguistic knowledge bases. In particular, we showed that two existing models for knowledge transfer based on descriptive attributes (Lampert *et al.*, 2009; Farhadi *et al.*, 2009) and distance measures (Fink, 2004; Bart and Ullman, 2005b) can be improved by determining potential sources and targets of knowledge transfer fully automatically, without the need for human supervision or additional training images. This improvement was achieved by exploiting the semantic relatedness of object classes and descriptive attributes, mined from a variety of linguistic knowledge bases, such as WordNet, Wikipedia, web search, and image search, using state-of-the-art semantic relatedness measures from the natural language processing (NLP) community (Zesch and Gurevych, 2010). Encouragingly, our experiments on the publicly available Animals-with-Attributes (AwA) (Lampert *et al.*, 2009) zero-shot recognition benchmark confirmed the recognition performance of our implementation to be en par with prior work using human supervision. Our experiments further provided valuable insights into the applicability of different linguistic knowledge bases for object class recognition, going beyond the limited use of NLP techniques in prior work.

9.2 FUTURE PERSPECTIVES

In this section, we highlight the current limitations of the work presented in this thesis, both on the level of individual contributions, and in the context of the bigger picture of object class recognition and the understanding of visual scenes. We further suggest directions for future work in order to overcome these limitations, and conclude with an outlook on potential future developments.

9.2.1 General object class recognition

Generalization to diverse object classes. The experiments conducted in connection with the object class models of chapters 5 (knowledge transfer experiments using three different quadruped classes) and 7 (shape learning from car 3D CAD models) are currently limited with respect to the object classes considered. While the presented experiments serve the purpose of validating the proposed methods for knowledge transfer, it would be desirable to explore their generalization to a more diverse set of object classes.

Unsupervised part discovery. The object class models presented in chapters 5 and 7 currently require manual part annotations, which identify corresponding parts across multiple object class training instances. While both can be provided using minimal labeling effort (point-and-click marking of distinct edge segments and connected mesh components, respectively), it is clearly desirable to replace or enrich manual supervision by unsupervised, data-driven part discovery. Data-driven part discovery can select parts with high discriminative power, tailored towards the specific task of recognition, at the cost of losing semantics (the most discriminative parts do not necessarily correspond to semantic parts).

Since both object class models have probabilistic formulations, the missing data problem (unknown part annotations) can be solved by Expectation Maximization (EM) Dempster *et al.* (1977), reminiscent of the constellation model (Fergus *et al.*, 2003). Since both models further use Markov Chain Monte Carlo (MCMC) sampling for inference, the specific variant of stochastic Expectation Maximization (Gilks *et al.*, 1996) applies, alternating MCMC inference and re-estimation of model parameters in a maximum likelihood framework.

Being trained from rendered 3D CAD models, the object class model of Chapter 7 lends itself to exploiting 3D surface mesh characteristics for part discovery, either as the sole source of information or in addition to rendered images. Part candidate sub-meshes could be obtained by applying part-aware mesh segmentation techniques from the computer graphics literature (Zhang *et al.*, 2005; Shalom *et al.*, 2008; Golovinskiy and Funkhouser, 2009; Liu *et al.*, 2009).

An alternative route to the discovery of corresponding parts has been pursued by Liebelt and Schmid (2010), although resulting in inferior recognition performance compared to our work (Chapter 7). Their approach approximates part correspondences by correspondences between cells of a regular grid, which is overlaid with the training meshes.

Partial occlusion handling. While the object class models of chapters 5 and 7 are to some extent robust to partial occlusion due to the combination of local part shape with a global spatial model, missing part evidence is currently not modeled explicitly. Both deliberately associate all model parts to a limited set of image features, even in cases where this association is unlikely according to the spatial model. One possible implementation of missing part evidence in the context of MCMC inference is given by trans-dimensional jump dynamics (Green, 1995), which allow to "jump" between Markov chain state spaces of varying dimensionality (object hypotheses with varying numbers of active parts). Although the tuning of reversible jump dynamics proved subtle in initial experiments, we expect the explicit modeling of missing part evidence to be beneficial for recognition performance.

Articulated pose modeling. As for partial occlusion, articulated pose is not modeled explicitly by the work presented in chapters 5 and 7. As a consequence, shape variations originating from pose variations can not be distinguished from inter-object class variations, reducing the discriminative power of the object class models. This limitation could be resolved by augmenting the search space for part position and scale by rotation, and adding corresponding terms to the likelihood function, as, e.g., proposed by Andriluka *et al.* (2009) in the context of a pictorial structures model (Felzenszwalb and Huttenlocher, 2000).

Multi-modal likelihoods. The probabilistic formulation of the object class models of chapters 5 and 7 currently chooses all constituent component likelihoods to be Gaussian, and thus uni-modal. For certain object classes, such as cars (Chapter 7), we would expect increased modeling accuracy from using multi-modal

densities, since class instances form clusters around multiple characteristic modes of appearance and geometry (e.g., sedan, convertible, SUV). The choice of Gaussian densities is motivated by history (the original constellation model (Fergus *et al.*, 2003) is Gaussian) rather than necessity. While the Gaussian assumption enables the combination of prior knowledge in the form of a covariance matrix with additional training examples in Chapter 5, the proposed DDMCMC inference technique does in principle not impose any restrictions on the functional form of involved densities (although the choice of density has an impact on mixing rates etc. (Gilks *et al.*, 1996)). Future work should take multi-modal densities into account, e.g., in the form of Gaussian mixture models (GMMs), or Gaussian Process Latent Variable Model (GPLVM) shape priors (Huang *et al.*, 2007).

Progressive proposals. The current implementation of DDMCMC inference in chapters 5 and 7 is using proposal distributions on the level of individual parts, resulting in a rather conservative exploration of the Markov chain state space, changing a single state component at each time step. It might be more efficient to change blocks of multiple components at a time (Gilks *et al.*, 1996), allowing for bigger "leaps" in the state space. Alternative proposal distributions could be generated from the marginal likelihoods of the components under consideration with respect to their spatial layout, or conditioned on the overall scale of the current hypothesis. In addition, non-data-driven diffusion moves could help exploring areas of the hypothesis space where local evidence is weak, and thus provides poor guidance to the search.

Multi-cue extension. Establishing unsupervised, data-driven part discovery for the object class models of chapters 5 and 7 could also facilitate their extension to multiple cues in addition to purely shape-based representations. From a pool of different available feature channels, part discovery could automatically choose the ones that best characterize the object class of interest, resulting in potentially heterogeneous object class models (Fergus *et al.*, 2004).

In the course of the multi-cue extension, the shading cues presented in Chapter 6 should be augmented by adding additional shading primitives, such as spheres (Nillius *et al.*, 2008). The shading cues would be further integrated into the probabilistic object class model formulation, replacing the preliminary late integration scheme of Chapter 6.

Probabilistic viewpoint estimation. While we demonstrate encouraging multi-view recognition performance for the object class model presented in Chapter 7 using a bank-of-detectors model (Thomas *et al.*, 2006), we expect improved performance for a proper integration of viewpoint estimation into the probabilistic formulation. While the bank-of-detectors is limited to a discrete set of trained viewpoints, a more accurate continuous viewpoint estimate could be incorporated as a latent variable into a single, viewpoint-aware formulation. This would further open the possibility to explicitly model intermediate viewpoints

by interpolation (Savarese and Fei-Fei, 2007) or morphing (Savarese and Fei-Fei, 2008). The resulting growth of the hypothesis space for recognition could be counterbalanced by adding proposals for the latent viewpoint variable, e.g., by training regressors on top of viewpoint-dependent part detector responses.

9.2.2 Knowledge transfer

Unsupervised knowledge transfer. In Chapter 5, transferable knowledge is manually specified, in the form of a subset of transferable parts and corresponding symmetry and spatial relations. On the one hand, this explicit transferability of individual model components constitutes an advantage over prior work in knowledge transfer (Fei-Fei *et al.*, 2006). On the other hand, it limits the applicability of the model to a few hand-picked cases, hindering scalability. The work on the automatic determination of potential sources and targets of knowledge transfer presented in Chapter 8 aims at overcoming exactly this limitation. However, in our experiments with various knowledge bases and semantic relatedness measures, the reliable extraction of precise knowledge for individual classes, such as part decomposition, proved difficult. As a consequence, we resorted to a less powerful object class representation, also in favor of comparability with previous work (Lampert *et al.*, 2009). Nevertheless, we consider the exploration of alternative means of harvesting knowledge an important ingredient of future work (Weikum and Theobald, 2010).

Joint learning from 3D CAD models and real world images. In Chapter 7, we argue that learning object class models from 3D CAD models alone, without using any real world images of the object class of interest, can outperform prior work using both Liebelt and Schmid (2010). However, we generally see the potential for further performance improvements by adding additional real world training images (since they are likely to reflect test image statistics better than images rendered from 3D CAD data), given an appropriate scheme for the alignment of both types of training data. A possible solution would use an initial model learned entirely from 3D CAD data to bootstrap the learning of a combined model, possibly in combination with data-driven, unsupervised part discovery.

Coupling between visual and language information. The attribute-based model for knowledge transfer adopted by Chapter 8 determines associations between object classes and descriptive attributes purely by their semantic relatedness using linguistic knowledge bases, without taking image information into account. As a consequence, the resulting visual attribute classifiers can not be guaranteed to truly reflect the intended semantics, but may learn incidentally correlated visual properties. This is particularly likely in case the attribute is not "visual" by itself (Boiy *et al.*, 2008). A more explicit connection between linguistic and image information could be achieved by extracting both types of information jointly, e.g., in Wikipedia or Flickr.

Automatic naming of learned visual properties. In Chapter 8, associations between object classes and attributes are determined by their semantic relatedness, inducing corresponding splits in the image training data, which are then manifested by training visual attribute classifiers on the basis of these splits. The discriminative power of the resulting attribute classifiers can hence not be guaranteed. An interesting alternative would be to reverse this process, by first determining a set of visual attribute classifiers with high discriminative power (e.g., by random training data splits (Farhadi *et al.*, 2009)), and naming the learned attributes afterwards, using linguistic knowledge bases. The implementation of this reverse lookup is, however, not obvious.

Enriching zero-shot recognition by additional training data. The automatic determination of potential sources and targets of knowledge transfer in Chapter 8 is specifically tailored towards zero-shot recognition, not using any training images of the object class of interest. It would be worth extending this approach by additionally exploiting available training data, e.g., based on the distance between known object classes and the object class of interest in some visual feature space (Fink, 2004; Bart and Ullman, 2005b,a).

Granularity and diversity of attribute modeling. In Chapter 8, attributes are associated to object classes. An alternative would be to perform the association using a finer granularity, such as on the level of individual images, or even image regions (Farhadi *et al.*, 2009; Wang and Forsyth, 2009), in order to improve their distinctiveness, and allow for more fine-grained predictions. Similarly, the proposed technique for automatically mining an inventory of attributes is currently limited to part attributes. It would obviously be desirable to increase the diversity of mined attributes to other properties as well (e.g., contextual attributes), in order to increase their descriptive power.

Functional reasoning. In our approach to recognizing objects according to functional categories in Chapter 4, we replace the representation of and reasoning over functions by purely visual learning from observation. While this provides a more robust alternative to earlier approaches based on functional reasoning (Stark and Bowyer, 1991; Stark *et al.*, 1993; Green *et al.*, 1995), it is greatly limited with respect to the functions that can be represented. Since the main limitation of early approaches consists in the gap between representational richness and the robust fitting of these representations to real world images, we believe there is large potential in combining the rich representations of early approaches with modern inference techniques, such as DDMCMC sampling. As demonstrated in the context of image parsing by Tu *et al.* (2005), DDMCMC allows to search spaces of extremely rich hypotheses, for which no obvious bottom-up recognition process exists. The key to success lies in the design of appropriate proposal distributions that guide the search, possibly relying on bottom-up processes governing individual hypothesis components (e.g., part detectors, as demonstrated in chapters 5 and 7).

Contextual reasoning. While this thesis explores the use of additional knowledge sources besides real world training images, such as linguistic knowledge bases and 3D CAD models, it does not consider the use of contextual knowledge at recognition time. Context has been shown to be potentially beneficial for recognition performance in a variety of incarnations (Wolf and Bileschi, 2006; Oliva and Torralba, 2007; Divvala *et al.*, 2009), ranging from local object support regions (Dalal and Triggs, 2005; Ramanan, 2007) to scene gist (Oliva and Torralba, 2001), coarse geometric scene layout (Hoiem *et al.*, 2006), and geographic location (Hays and Efros, 2008).

9.2.3 The bigger picture

While the previous sections are concerned with limitations and corresponding future work items related to the individual contributions of this thesis, this section tries to give an outlook on object class recognition and knowledge transfer taking a broader perspective, anticipating measures on a larger scale in order to come closer to the ultimate goal of obtaining a human-like understanding of unrestricted visual scenes, or at least recognizing large quantities of object classes simultaneously with sufficient accuracy.

Richness in representation. Rich representations in terms of multiple available low-level feature channels have been accepted as a promising route to achieving high recognition performance on a limited set of object classes (Everingham *et al.*, 2010). We believe that providing a rich inventory of available building blocks also from higher-level visual representations constitutes the basis of a successful recognition of multiple classes, and, ultimately, the detailed interpretation of visual scenes. It seems reasonable to assume that different object classes are best modeled by different object class representations, and that a large supply of available representations increases the likelihood of having a good fit available. Following this argumentation, the supply of representations should comprise both parametric and non-parametric, representative and discriminative models, models for specific object classes and generic attributes, top-down searches and bottom-up groupings and segmentors, in order to benefit from the strengths of the individual representations while being able to compensate their weaknesses.

Integration of knowledge sources. A second ingredient of scalable and powerful recognition is arguably the integration of knowledge from all available, probably heterogeneous sources. It seems unreasonable to impose artificial restrictions on the kinds of allowed training data. On the contrary, all available knowledge, image training data, 3D CAD data, linguistic knowledge bases, etc. should be combined to aid recognition, allowing to profit from their complementary nature. On the other hand, integrating different knowledge sources poses certain challenges, both with respect to their differing syntax and semantics, and the increased computational cost of maintaining the combined

knowledge. A first challenge consists in establishing robust correspondences between entities from different sources, explicitly taking into account the uncertainty inherent to all made associations, and the differing reliability of different knowledge sources. A second challenge arises from the increased complexity of the joint modeling of multiple sources with respect to memory and computation time, motivating a shift of focus from powerful learning techniques back to efficient storage structures, indexing schemes, and optimal approximate query matching.

Explorative scene interpretation. Even given the rapid advancement of modern computing machinery, maintaining a separate, pre-trained model for each of tens of thousands of object classes, a plethora of generic attributes, and further discriminative combinations of both (such as grouplets (Yao and Fei-Fei, 2010)) for the interpretation of a single previously unseen image still seems unrealistic. Instead, it seems more likely that a clever combination of pre-computed models and on-demand processing can prove successful. Having a large inventory of representations and potential knowledge sources available will require the careful planning of costly actions involved in the recognition process, including conditional branching and backtracking to earlier stages, in the spirit of an explorative analysis [14].

As a simplified example, consider the interpretation of an unrestricted, previously unseen visual scene. Since the gist (Oliva and Torralba, 2001) of an image can reveal important contextual information about the general kind of visual scene that is pictured, it is reasonable to query a community photo sharing service for similar images, using non-parametric methods. The returned query results will typically contain noisy tags, hinting on their contents. On the basis of these tags, using semantic relatedness measures, links between general kinds of visual scenes and object classes to be expected in the context of these scenes can be established, and corresponding object class detectors triggered on visually salient image regions. All regions that can be explained with sufficient confidence by the detectors can be removed from further consideration (or kept for computing alternative interpretations). All unexplained regions of sufficient saliency can be further analyzed by applying perceptual grouping techniques, revealing underlying structures, and characterized by generic attribute and simile classifiers, to yield an as precise as possible interpretation of the scene.

[14] http://www.semantic-robot-vision-challenge.org/

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

W.-K. Ahn and W. Brewer (1993). Psychological Studies of Explanation-Based Learning, in G. DeJong (ed.), *Investigating Explanation-Based Learning 1993*, Kluwer. 74

Y. Amit, M. Fink, N. Srebro, and S. Ullman (2007). Uncovering shared structures in multiclass classification, in *Proceedings of International Conference on Machine learning 2007*. 34, 44, 75

M. Andriluka, S. Roth, and B. Schiele (2008). People-Tracking-by-Detection and People-Detection-by-Tracking, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08) 2008*. 21

M. Andriluka, S. Roth, and B. Schiele (2009). Pictorial structures revisited: People detection and articulated pose estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. 22, 33, 112, 140, 142

M. Arie-Nachimson and R. Basri (2009). Constructing Implicit 3D Shape Models for Pose Estimation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. 25, 33, 45, 109, 113

A. Bar-Hillel and D. Weinshall (2008). Efficient Learning of Relational Object Class Models, *International Journal of Computer Vision*, vol. 77(1-3), pp. 175–198. 37

K. Barnard, P. Duygulu, D. A. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan (2003). Matching Words and Pictures, *Journal of Machine Learning Research*. 39, 124

K. Barnard and K. Yanai (2006). Mutual information of words and pictures, in *Information Theory and Applications 2006*. 41, 124

E. Bart and S. Ullman (2005a). Cross-Generalization: Learning Novel Classes from a Single Example by Feature Replacement, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. 35, 44, 75, 124, 145

E. Bart and S. Ullman (2005b). Single-example learning of novel classes using representation by similarity, in *Proceedings of the British Machine Vision Conference (BMVC) 2005*. 2, 3, 35, 44, 45, 74, 75, 122, 124, 137, 141, 145

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool (2008). Speeded-Up Robust Features (SURF), *Computer Vision and Image Understanding*, vol. 110(3), pp. 346–359. 27, 33, 36

S. Belongie, J. Malik, and J. Puzicha (2000). Shape Context: A New Descriptor for Shape Matching and Object Recognition, in *Advances in Neural Information Processing Systems (NIPS) 2000*. 16, 17, 31, 48, 50, 52, 112, 140

S. Belongie, J. Malik, and J. Puzicha (2001). *Shape matching and object recognition using shape contexts*. 19, 22, 33

S. Ben-David and R. Schuller (2003). Exploiting Task Relatedness for Multiple Task Learning, in *COLT 2003*. 3, 34, 44, 75

A. C. Berg, T. L. Berg, and J. Malik (2005). Shape Matching and Object Recognition Using Low Distortion Correspondences, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. 51

A. C. Berg and J. Malik (2001). Geometric Blur for Template Matching, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2001*. 31, 48, 50, 51

M. Berland and E. Charniak (1999). Finding parts in very large corpora, in *Annual Meeting of the ACL 1999*. 45, 129

I. Biedermann (1987). Recognition-by-components: a theory of human image understanding, *Psychological Review*, vol. 94(2), pp. 115–147. 95

L. Bogoni and R. Bajcsy (1995). Interactive Recognition and Representation of Functionality, *Computer Vision and Image Understanding*, vol. 62(2), pp. 194–214. 42, 44, 64, 140

E. Boiy, K. Deschacht, and M.-F. Moens (2008). Learning Visual Entities and Their Visual Attributes from Text Corpora, in *DEXA 2008*. 40, 41, 124, 144

A. Bosch, A. Zisserman, and X. Munoz (2007). Representing shape with a spatial pyramid kernel, in *CIVR 2007*. 33, 36, 43

Y. Boykov and V. Kolmogorov (2004). An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26(9), pp. 1124–1137. 32

M. Brady and H. Asada (1984). Smoothed local symmetries and their implementation, in *The International Journal of Robotics Research 1984*. 20, 32, 76, 78, 139

J. E. Bresenham (1965). Algorithm for computer control of a digital plotter, *IBM Systems Journal*. 98

R. Brooks (1983). Model-based three dimensional interpretation of two dimensional images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 76

R. Brooks, R. Creiner, and T. Binford (1979). The ACRONYM Model-Based Vision System, in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 1979*. 1, 20, 25, 44, 108, 110, 112

A. Budanitsky and G. Hirst (2006). Evaluating WordNet-based measures of lexical semantic relatedness, *Computational Linguistics*. 45, 128

M. C. Burl and P. Perona (1996). Recognition of Planar Object Classes, in *CVPR '96: Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96) 1996*. 23

M. C. Burl, M. Weber, and P. Perona (1998). A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry, in *Proceedings of the European Conference on Computer Vision (ECCV) 1998*. 23, 32

C.-C. Chang and C.-J. Lin (2001). *LIBSVM: a library for support vector machines*. 130

T. Cootes (2000). *An Introduction to Active Shape Models*. 18, 77, 79, 114

G. Csurka, C. Dance, L. Fan, J. Willarnowski, and C. Bray (2004). Visual Categorization with Bags of Keypoints, in *Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision 2004*. 17, 33, 48, 63, 93

N. Dalal and B. Triggs (2005). Histograms of Oriented Gradients for Human Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. 23, 146

B. Delezoide, G. Pitel, H. L. Borgne, G. Greffenstette, P.-A. Moëllic, and C. Millet (2008). Object/Background Scene Classification in Photographs Using Linguistic Statistics from the Web, in *OntoImage 2008*. 40, 124

A. P. Dempster, N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B*, vol. 39, pp. 1–38. 23, 26, 27, 28, 39, 142

S. J. Dickinson, A. P. Pentland, and A. Rosenfeld (1992). From volumes to views: an approach to 3-D object recognition, *CVGIP: Image Underst.*, vol. 55(2), pp. 130–154. 42, 44

S. K. Divvala, D. Hoiem, J. H. Hays, A. Efros, and M. Hebert (2009). An Empirical Study of Context in Object Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. 146

J. Dorsey, H. Rushmeier, and F. X. Sillion (2007). *Digital Modeling of Material Appearance*, Morgan Kaufmann. 96

M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman (2010). The Pascal Visual Object Classes (VOC) Challenge, *International Journal of Computer Vision*, vol. 88(2), pp. 303–338. 1, 63, 108, 137, 146

A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth (2009). Describing objects by their attributes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. 3, 9, 36, 37, 43, 45, 123, 137, 141, 145

L. Fei-Fei, R. Fergus, and P. Perona (2004). Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories, in *CVPR, Workshop on Generative Model Based Vision 2004*. 48, 50

L. Fei-Fei, R. Fergus, and P. Perona (2006). One-Shot Learning of Object Categories, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28(4), pp. 594–611. 24, 38, 44, 74, 76, 124, 137, 140, 144

C. Fellbaum (1998). *WordNet: An Electronical Lexical Database*, MIT Press, Cambridge, MA. 38, 123, 128

P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan (2009). Object Detection with Discriminatively Trained Part Based Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1, 19, 23

P. F. Felzenszwalb and D. P. Huttenlocher (2000). Efficient Matching of Pictorial Structures, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. 22, 142

A. Ferencz, E. Learned-Miller, and J. Malik (2005). Building a Classification Cascade for Visual Identification from One Example, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2005*. 74

R. Fergus, P. Perona, and A. Zisserman (2003). Object Class Recognition by Unsupervised Scale-Invariant Learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2003*. 1, 2, 12, 17, 23, 24, 26, 32, 38, 39, 48, 76, 78, 93, 110, 113, 139, 142, 143

R. Fergus, P. Perona, and A. Zisserman (2004). A Visual Category Filter for Google Images, in *Proceedings of the European Conference on Computer Vision (ECCV) 2004*. 24, 143

R. Fergus, M. Weber, and P. Perona (2001). Efficient Methods for Object Recognition using the Constellation Model, Technical report, California Institute of Technology. 23, 32

V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid (2006a). Groups of Adjacent Contour Segments for Object Detection, *Rapport De Recherche Inria*. 47, 48, 49, 50, 66

V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid (2008). Groups of Adjacent Contour Segments for Object Detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30(1), pp. 36–51. 18, 31, 32, 43, 63, 139, 140

V. Ferrari, F. Jurie, and C. Schmid (2007). Accurate Object Detection with Deformable Shape Models Learnt from Images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. 18, 19, 32, 77, 80, 81, 93, 102, 139

V. Ferrari, T. Tuytelaars, and L. J. V. Gool (2006b). Object Detection by Contour Segment Networks, in *Proceedings of the European Conference on Computer Vision (ECCV) 2006*. 18, 32, 43, 50, 54, 66, 69, 77, 80, 95, 98, 102, 139, 140

V. Ferrari, T. Tuytelaars, and L. V. Gool (2004). Integrating Multiple Model Views for Object Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004*. 25

V. Ferrari and A. Zisserman (2007). Learning Visual Attributes, in *Advances in Neural Information Processing Systems (NIPS) 2007*. 35, 45, 123

S. Fidler, M. Boben, and A. Leonardis (2009). Evaluating multi-class learning strategies in a generative hierarchical framework for object detection, in *Advances in Neural Information Processing Systems (NIPS) 2009*. 139

M. Fink (2004). Object Classification from a Single Example Utilizing Class Relevance Pseudo-Metrics, in *Advances in Neural Information Processing Systems (NIPS) 2004*. 2, 3, 34, 35, 44, 45, 74, 75, 122, 124, 137, 141, 145

M. Fink and S. Ullman (2008). From Aardvark to Zorro: A Benchmark for Mammal Image Classification, *International Journal of Computer Vision*, vol. 77(1-3), pp. 143–156. 34, 84

M. A. Fischler and R. C. Bolles (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the* ACM, vol. 24(6), pp. 381–395. 25, 32, 98

M. A. Fischler and R. A. Elschlager (1973). The Representation and Matching of Pictorial Structures, IEEE *Transactions on Computers*, vol. 22(1), pp. 67–92. 22, 110

W. T. Freeman and E. H. Adelson (1991). The Design and Use of Steerable Filters, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13(9), pp. 891–906. 16

Y. Freund and R. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*. 22, 34, 37, 112, 140

M. Fritz, M. Andriluka, S. Fidler, M. Stark, A. Leonardis, and B. Schiele (2010). *Categorical Perception*, chapter Categorical Perception, no. 8 in Cognitive Systems Monographs, Springer. 11

M. Fritz and B. Schiele (2008). Decomposition, Discovery and Detection of Visual Categories Using Topic Models, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. 19, 32, 80, 81, 139

E. Gabrilovich and S. Markovitch (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2007*. 45, 128

D. Gavrila (2000). Pedestrian Detection from a Moving Vehicle, in *Proceedings of the European Conference on Computer Vision (ECCV) 2000*. 17

J. J. Gibson (1977). The Theory of Affordance, in *Percieving, Acting, and Knowing 1977*, Lawrence Erlbaum Associates, Hillsdale, NJ. 34, 64, 138

J. J. Gibson (1979). *The Ecological Approach to Visual Perception*, Houghton Mifflin. 1

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo In Practice*, Chapman & Hall/CRC. 30, 79, 114, 142, 143

G. Gill and M. Levine (2009). Multi-view Object Detection Based on Spatial Consistency in a Low Dimensional Space, in *Pattern Recognition, Proceedings of the DAGM Symposium 2009*. 27, 28, 44, 45, 109, 116, 117, 140

A. Golovinskiy and T. Funkhouser (2009). Consistent Segmentation of 3D Models, *Computers and Graphics (Shape Modeling International 09)*, vol. 33(3), pp. 262–269. 142

L. J. V. Gool, T. Moons, and D. Ungureanu (1996). Affine/ Photometric Invariants for Planar Intensity Patterns, in *Proceedings of the European Conference on Computer Vision (ECCV) 1996*. 16

K. Green, D. Eggert, L. Stark, and K. Bowyer (1995). Generic Recognition of Articulated Objects through Reasoning about Potential Function, *Computer Vision and Image Understanding*, vol. 62(2), pp. 177–193. 41, 43, 140, 145

P. J. Green (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, vol. 82, pp. 711–732. 9, 30, 142

G. Griffin, A. Holub, and P. Perona (2007). Caltech-256 Object Category Dataset, Technical report 7694, California Institute of Technology. 63

J. Haddon and D. Forsyth (1998). Shape representations from shading primitives, in *Proceedings of the European Conference on Computer Vision (ECCV) 1998*. 28, 29, 32, 95, 96, 139

R. M. Haralick, C.-N. Lee, K. Ottenberg, and M. Nölle (1994). Review and analysis of solutions of the three point perspective pose estimation problem, *Int. J. Comput. Vision*, vol. 13(3), pp. 331–356. 27

C. Harris and M. J. Stephens (1988). A Combined Corner and Edge Detector, in *Alvey Conference 1988*. 16, 52

N. Hawes, M. Zillich, and J. Wyatt (2007). BALT & CAST: Middleware for Cognitive Robotics, in *Proceedings of IEEE RO-MAN 2007 2007*. 12

J. Hays and A. Efros (2008). IM2GPS: estimating geographic information from a single image, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. 146

S. Helmer and D. G. Lowe (2004). Object Class Recognition with Many Local Features, in *CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 12 2004*. 23, 32

A. Hertzmann (1999). Introduction to 3D Non-Photorealistic Rendering: Silhouettes and Outlines, in *SIGGRAPH 99 Course Notes 1999*. 112

D. Hoiem, A. A. Efros, and M. Hebert (2006). Putting Objects in Perspective, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2006*. 146

B. Horn and M. Brooks (1989). *Shape from Shading*, MIT Press. 28, 94

Y. Huang, Q. Liu, and D. N. Metaxas (2007). A Component Based Deformable Model for Generalized Face Alignment, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. 143

M. Jamieson, A. Fazly, S. Stevenson, S. Dickinson, and S. Wachsmuth (2010). Using Language to Learn Structured Appearance Models for Image Annotation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32(1), pp. 148–164. 4, 39, 45

M. J. Jones and J. M. Rehg (1999). Statistical Color Models with Application to Skin Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1999*. 32, 66

F. Jurie and C. Schmid (2004). Scale-invariant shape features for recognition of object categories, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004*. 84

T. Kadir, A. Zisserman, and M. Brady (2004). An Affine Invariant Salient Region Detector, in *Proceedings of the European Conference on Computer Vision (ECCV) 2004*. 16, 17, 23, 50, 52

E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann (2009). Image Sequence Geolocation with Human Travel Priors, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. 40

Y. Ke and R. Sukthankar (2004). PCA-SIFT: A More Distinctive Representation for Local Image Descriptors, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004*. 16, 49

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda (2006). Learning Systems of Concepts with an Infinite Relational Model, in *AAAI Conference on Artificial Intelligence 2006*. 126, 130

S. M. Khan, P. Yan, and M. Shah (2007). A homographic framework for the fusion of multi-view silhouettes, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. 26

A. Kilgarriff and G. Grefenstette (2003). Introduction to the Special Issue on the Web as Corpus, *Computational Linguistics*, vol. 29, pp. 333–347. 45, 129

H. Kjellström, J. Romero, D. Martínez, and D. Kragić (2008). Simultaneous Visual Recognition of Manipulation Actions and Manipulated Objects, in *Proceedings of the European Conference on Computer Vision (ECCV) 2008*. 4, 42, 44

D. A. Kleffner and V. S. Ramachandran (1992). On the perception of shape from shading, *Perception and Psychophysics*, vol. 52, pp. 18–36. 28, 93

J. Koenderink and A. van Doorn (1992). Surface Shape and Curvature Scales, *Image and Vision Computing*. 95

J. Koenderink, A. Van Doorn, C. Christou, and J. Lappin (1996). Perturbation study of shading in pictures, *Perception*. 28, 93

J. J. Koenderink and A. J. van Doom (1987). Representation of local geometry in the visual system, *Biol. Cybern.*, vol. 55(6), pp. 367–375. 16

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar (2009). Attribute and Simile Classifiers for Face Verification, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. 3, 37, 43, 45, 123

A. Kushal, C. Schmid, and J. Ponce (2007). Flexible Object Models for Category-Level 3D Object Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. 26

J. Lafferty, A. McCallum, and F. Pereira (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in *Proceedings of International Conference on Machine learning 2001*. 32, 42, 65

C. H. Lampert, H. Nickisch, and S. Harmeling (2009). Learning to detect unseen object classes by between-class attribute transfer, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. 3, 8, 13, 33, 36, 43, 45, 122, 123, 124, 125, 126, 130, 131, 132, 133, 135, 137, 141, 144

S. Lazebnik, C. Schmid, and J. Ponce (2005). A Sparse Texture Representation Using Local Affine Regions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(8), pp. 1265–1278. 16

S. Lazebnik, C. Schmid, and J. Ponce (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2006*. 22, 28, 63

M. W. Lee and I. Cohen (2004). Proposal Maps Driven MCMC for Estimating Human Body Pose in Static Images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004*. 31, 32, 79

B. Leibe, A. Leonardis, and B. Schiele (2006a). An Implicit Shape Model for Combined Object Categorization and Segmentation, in *Toward Category-Level Object Recognition 2006*. 1, 17, 18, 21, 25, 26, 27, 31, 45, 68, 76, 110, 140

B. Leibe, K. Mikolajczyk, and B. Schiele (2006b). Efficient Clustering and Matching for Object Class Recognition, in *Proceedings of the British Machine Vision Conference (BMVC) 2006*. 53

B. Leibe and B. Schiele (2003). Analyzing Appearance and Contour Based Methods for Object Categorization, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2003*. 49

B. Leibe, E. Seemann, and B. Schiele (2005). Pedestrian Detection in Crowded Scenes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. 17

M. Leordeanu, M. Hebert, and R. Sukthankar (2007). Beyond Local Appearance: Category Recognition from Pairwise Interactions of Simple Features, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. 32, 76

K. Levenberg (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares, *The Quarterly of Applied Mathematics*. 98

K. Levi, M. Fink, and Y. Weiss (2004). Learning From a Small Number of Training Examples by Exploiting Object Categories, in *Workshop of Learning in Computer Vision 2004*. 37, 74, 75

L. J. Li, R. Socher, and L. F. Fei (2009). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. 4, 39, 45, 124

J. Lichtenauer, E. Hendriks, and M. Reinders (2005). Isophote Properties as Features for Object Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. 29, 33, 95

J. Liebelt and C. Schmid (2010). Multi-View Object Class Detection with a 3D Geometric Model, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. 4, 27, 44, 45, 109, 110, 111, 116, 117, 118, 140, 142, 144

J. Liebelt, C. Schmid, and K. Schertler (2008). Viewpoint-independent object class detection using 3D Feature Maps, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. 4, 27, 33, 44, 110, 111, 113

D. Lin (1998). An Information-Theoretic Definition of Similarity, in *Proceedings of International Conference on Machine learning 1998*. 45, 128

R. Liu, H. Zhang, A. Shamir, and D. Cohen-Or (2009). A Part-Aware Surface Metric for Shape Analysis, *Computer Graphics Forum (Special Issue of Eurographics 2009)*, vol. 28(2), pp. 397–406. 142

Y. Liu, J. H. Hays, Y.-Q. Xu, and H.-Y. Shum (2005). Digital Papercutting, in *Computer Graphics Proceedings, Annual Conference Series, (ACM SIGGRAPH) 2005*. 20

D. Lowe (1987). Three-Dimensional Object Recognition from Single Two-Dimensional Images, *Artificial Intelligence*, vol. 31, pp. 355–395. 1, 25, 44, 108, 110, 112

D. Lowe (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, vol. 60(2), pp. 91–110. 16, 17, 26, 27, 33, 36, 48, 50, 52, 55, 66, 93, 108

G. Loy and J. olof Eklundh (2006). Detecting symmetry and symmetric constellations of features, in *Proceedings of the European Conference on Computer Vision (ECCV) 2006*. 20

S. Maji and J. Malik (2009). Object Detection using a Max-Margin Hough Transform, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. 19, 21, 22, 81, 82

D. Marr and H. Nishihara (1978). Representation and recognition of the spatial organization of three-dimensional shapes, *Proc. Roy. Soc. London B 200*, pp. 269–194. 1, 24, 25, 44, 108, 110, 112

M. Marszalek and C. Schmid (2007). Semantic Hierarchies for Visual Object Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. 2, 3, 38, 124

D. R. Martin, C. Fowlkes, and J. Malik (2004). Learning to Detect Natural Image Boundaries Using Local Brightness, Color, and Texture Cues, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26(5), pp. 530–549. 18, 51, 77

J. Matas, O. Chum, M. Urban, and T. Pajdla (2002). Robust Wide Baseline Stereo from Maximally Stable Extremal Regions, in *Proceedings of the British Machine Vision Conference (BMVC) 2002*. 16

A. McCallum and K. Nigam (1998). A comparison of event models for Naive Bayes text classification, in *AAAI, Workshop on Learning for Text Categorization 1998*. 54

K. Mikolajczyk, B. Leibe, and B. Schiele (2005a). Local Features for Object Class Recognition, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2005*. 16, 17, 31, 49, 53, 54, 55, 60

K. Mikolajczyk, B. Leibe, and B. Schiele (2006). Multiple Object Class Detection with a Generative Model, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2006*. 17, 48

K. Mikolajczyk and C. Schmid (2002). An Affine Invariant Interest Point Detector, in *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I 2002*. 16, 17

K. Mikolajczyk and C. Schmid (2004). Scale & Affine Invariant Interest Point Detectors, *International Journal of Computer Vision*, vol. 60(1), pp. 63–86. 50

K. Mikolajczyk and C. Schmid (2005). A Performance Evaluation of Local Descriptors., *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(10), pp. 1615–1630. 16, 17, 48, 49, 50, 52

K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. J. V. Gool (2005b). A Comparison of Affine Region Detectors, *International Journal of Computer Vision*. 16, 17, 49, 50, 52, 66

E. Miller, N. Matsakis, and P. Viola (2000). Learning from One Example Through Shared Densities on Transforms, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2000*. 38, 44, 74

G. A. Miller (1995). WordNet: a lexical database for English, *Communications of the* ACM, vol. 38(11), pp. 39–41. 40

P. Moreels and P. Perona (2005). Evaluation of Features Detectors and Descriptors Based on 3D Objects, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2005*. 17, 49

G. Mori, X. Ren, A. Efros, and J. Malik (2004). Recovering Human Body Configurations: Combining Segmentation and Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004*. 29, 32, 95

M. Morrone and D. Burr (1988). Feature detection in human vision: a phase dependent energy model, *Proc. Royal Soc. London Bulletin*, pp. 221–245. 51

Y. Moses, S. Ullman, and S. Edelman (1993). Generalization across changes in Illumination and Viewing Position in Upright and Inverted Faces, Technical report, Weizmann Institute. 74

J. L. Mundy (2006). Object Recognition in the Geometric Era: A Retrospective., in *Toward Category-Level Object Recognition 2006*. 47

K. P. Murphy, Y. Weiss, and M. I. Jordan (1999). Loopy Belief Propagation for Approximate Inference: An Empirical Study, in *Conference on Uncertainty in Artificial Intelligence 1999*. 27

S. A. Nene, S. K. Nayar, and H. Murase (1996). Columbia Object Image Library (COIL-20), *Technical Report CUCS-006-96*. 29, 95

R. Nevatia and T. Binford (1977). Description and Recognition of Curved Objects, *Artificial Intelligence*, vol. 8, pp. 77–98. 1, 24, 44, 76, 108, 110, 112

A. Niculescu-Mizil and R. Caruana (2005). Obtaining Calibrated Probabilities from Boosting, in *Conference on Uncertainty in Artificial Intelligence 2005*. 113

P. Nillius, J. Sullivan, and A. Argyros (2008). Shadiopen relng Models for Illumination and Reflectance Invariant Shape Detectors, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. 29, 32, 33, 44, 95, 143

A. Oliva and A. Torralba (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, *International Journal of Computer Vision*, vol. 42(3), pp. 145–175. 146, 147

A. Oliva and A. Torralba (2007). The role of context in object recognition, *Trends in Cognitive Sciences*, vol. 11(12). 146

B. Ommer and J. Malik (2009). Multi-scale object detection by clustering lines, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. 22, 81, 82

A. Opelt, A. Pinz, and A. Zisserman (2006). A Boundary-Fragment-Model for Object Detection, in *Proceedings of the European Conference on Computer Vision (ECCV) 2006*. 47

D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith (1991). Default Probability, *Cognitive Science*, vol. 15(2), pp. 251–269. 36, 45, 126, 130

M. Ozuysal, V. Lepetit, and P. Fua (2009). Pose estimation for category specific multiview object localization, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. 109

M. Park, S. Lee, P.-C. Chen, S. Kashyap, A. A. Butt, and Y. Liu (2008). Performance Evaluation of State-of-the-Art Discrete Symmetry Detection Algorithms, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. 20, 21, 32, 76

A. Pentland (1986). Perceptual Organization and the Representation of Natural Form, *Artificial Intelligence*, vol. 28, pp. 293–331. 1, 19, 42, 44, 108, 110, 112

A. Popescu, C. Millet, and P.-A. Moëllic (2007). Ontology driven content based image retrieval, in *CIVR 2007*. 40, 45, 124

V. S. N. Prasad and L. S. Davis (2005). Detecting Rotational Symmetries, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2005*. 20

D. Ramanan (2007). Using Segmentation to Verify Object Hypotheses, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. 146

S. Ravishankar, A. Jain, and A. Mittal (2008). Multi-stage Contour Based Detection of Deformable Objects, in *Proceedings of the European Conference on Computer Vision (ECCV) 2008*. 19, 81

P. Resnik (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy, in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 1995*. 128

E. Rivlin, S. J. Dickinson, and A. Rosenfeld (1995). Recognition by Functional Parts, *Computer Vision and Image Understanding*, vol. 62(2), pp. 164–176. 42, 43, 64, 140

M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele (2010). What Helps Where – And Why? Semantic Relatedness for Knowledge Transfer, in *CVPR 2010*. 13

E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. B. Braem (1976). Basic objects in natural categories, *Cognitive Psychology*. 4, 33, 41, 63, 138

P. Saint-Marc, H. Rom, and G. Medioni (1993). B-spline Contour Representation and Symmetry Detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15(11), pp. 1191–97. 20, 32, 76, 78

S. Savarese and L. Fei-Fei (2007). 3D generic object categorization, localization and pose estimation., in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. 24, 25, 26, 45, 115, 140, 144

S. Savarese and L. Fei-Fei (2008). View synthesis for recognizing unseen poses of object classes., in *IEEE European Conference on Computer Vision. 2008*. 24, 26, 45, 144

A. Saxena, J. Driemeyer, and A. Y. Ng (2007). Robotic Grasping of Novel Objects using Vision, *IJRR*. 42, 43, 64

F. Schaffalitzky and A. Zisserman (2002). Multi-view Matching for Unordered Image Sets, or "How Do I Organize My Holiday Snaps?", in *Proceedings of the European Conference on Computer Vision (ECCV) 2002*. 16

F. Schaffalitzky and A. Zisserman (2003). Automated location matching in movies, *Computer Vision and Image Understanding*, vol. 92(2-3), pp. 236–264. 16

C. Schmid and R. Mohr (1997). Local Grayvalue Invariants for Image Retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19(5), pp. 530–535. 49

E. Seemann, M. Fritz, and B. Schiele (2007). Towards Robust Pedestrian Detection in Crowded Image Sequences, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. 21

E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele (2005). An Evaluation of Local Shape-Based Features for Pedestrian Detection, in *Proceedings of the British Machine Vision Conference (BMVC) 2005*. 17, 49

S. Shalom, L. Shapira, A. Shamir, and D. Cohen-Or (2008). Part Analogies in Sets of Objects, in *Proceedings of Eurographics Symposium on 3D Object Retrieval 2008*. 111, 142

E. Shechtman and M. Irani (2007). Matching Local Self-Similarities across Images and Videos, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. 33, 36

D. Skocaj, M. Kristan, A. Vrecko, A. Leonardis, M. Fritz, M. Stark, B. Schiele, S. Hongeng, and J. L. Wyatt (2010). *Multi-Modal Learning*, chapter Multi-Modal Learning, no. 8 in Cognitive Systems Monographs, Springer. 12

P. Srinivasan, Q. Zhu, and J. Shi (2010). Many-to-one Contour Matching for Describing and Discriminating Object Shape, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. 19, 81, 82

L. Stark and K. Bowyer (1991). Achieving Generalized Object Recognition through Reasoning about Association of Function to Structure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13(10), pp. 1097–1104. 4, 41, 42, 43, 64, 140, 145

L. Stark, A. Hoover, D. Goldgof, and K. Bowyer (1993). Function-Based Recognition from Incomplete Knowledge of Shape, in *WQV93 1993*. 41, 43, 64, 140, 145

M. Stark, M. Goesele, and B. Schiele (2009a). Shading Cues for Object Class Detection, in *2nd International IEEE Workshop on 3D Representation for Recognition (3dRR-09) 2009*. 13

M. Stark, M. Goesele, and B. Schiele (2009b). A Shape-Based Object Class Model for Knowledge Transfer, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. 12

M. Stark, M. Goesele, and B. Schiele (2010). Back to the Future: Learning Shape Models from 3D CAD Data, in *Proceedings of the British Machine Vision Conference (BMVC) 2010*. 13

M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele (2008). Functional Object Class Detection Based on Learned Affordance Cues, in *Proceedings of the International Conference on Computer Vision Systems (ICVS) 2008*. 12

M. Stark and B. Schiele (2007). How Good are Local Features for Classes of Geometric Objects, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. 11, 55

H. Su, M. Sun, L. Fei-Fei, and S. Savarese (2009). Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories., in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. 24, 26, 27, 28, 33, 44, 45, 109, 113, 116, 117, 140

E. Sudderth, A. Torralba, W. T. Freeman, and A. Willsky (2008). Describing visual scenes using transformed objects and parts, *International Journal of Computer Vision*. 44, 76

J. Sun, W. W. Zhang, X. Tang, and H. Y. Shum (2006). Background Cut, in *Proceedings of the European Conference on Computer Vision (ECCV) 2006*. 32, 65

M. Sun, H. Su, S. Savarese, and L. Fei-Fei (2009). A Multi-View Probabilistic Model for 3D Object Classes, in *Proc. Computer Vision and Pattern Recognition 2009*. 24, 26, 45

A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool (2006). Towards Multi-View Object Class Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2006*. 25, 26, 33, 109, 112, 113, 139, 143

S. Thrun (1996). Is Learning The n-th Thing Any Easier Than Learning the First, in *Advances in Neural Information Processing Systems (NIPS) 1996*. 3, 35, 75, 124

E. Tola, V. Lepetit, and P. Fua (2010). DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 815–830. 28

M. Torki and A. Elgammal (2010). Putting Local Features on a Manifold, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. 139

A. Torralba, K. Murphy, and W. Freeman (2004). Sharing visual features for multiclass and multiview object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004*. 2, 3, 34, 37, 44, 54, 75, 137

Z. Tu, X. Chen, A. Yuille, and S. Zhu (2005). Image Parsing: Unifying Segmentation, Detection And Recognition, *International Journal of Computer Vision*. 30, 32, 79, 114, 145

T. Tuytelaars and L. J. V. Gool (1999). Content-Based Image Retrieval Based on Local Affinely Invariant Regions, in *VISUAL '99: Proceedings of the Third International Conference on Visual Information and Information Systems 1999*. 16

T. Tuytelaars and L. V. Gool (2000). Wide Baseline Stereo Matching based on Local, Affinely Invariant Regions, in *Proceedings of the British Machine Vision Conference (BMVC) 2000*. 16

K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek (2010). Evaluating Color Descriptors for Object and Scene Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (in press). 33, 36

P. A. Viola and M. J. Jones (2001). Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade, in *Advances in Neural Information Processing Systems (NIPS) 2001*. 30

G. Wang and D. Forsyth (2009). Joint Learning of Visual Attributes, Object Classes and Visual Saliency, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. 3, 37, 123, 145

H. Wang, X. Jiang, L.-T. Chia, and A.-H. Tan (2008). Ontology enhanced web image retrieval: aided by wikipedia & spreading activation theory, in *MIR 2008*. 40, 124

M. Weber, M. Welling, and P. Perona (2000). Unsupervised learning of models for recognition, in *Proceedings of the European Conference on Computer Vision (ECCV) 2000*. 2, 23, 26, 139

G. Weikum and M. Theobald (2010). From information to knowledge: harvesting entities and relationships from web sources, in *PODS '10: Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems of data 2010*. 144

D. Weinshall (1992). Local shape approximation from shading, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1992*. 28, 32, 95, 139

P. H. Winston, B. Katz, T. O. Binford, and M. R. Lowry (1983). Learning Physical Descriptions From Functional Definitions, Examples, and Precedents, in *AAAI Conference on Artificial Intelligence 1983*. 4, 43, 64

C. Wojek, S. Roth, K. Schindler, and B. Schiele (2010). Monocular 3D Scene Modeling and Inference: Understanding Multi-Object Traffic Scenes, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. 30, 32

L. Wolf and S. Bileschi (2006). A Critical View of Context, *International Journal of Computer Vision*, vol. 69(2). 146

P. L. Worthington and E. R. Hancock (2001). Object Recognition Using Shape-from-Shading, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(5), pp. 535–542. 29, 95

J. Wu, W. Smith, and E. Hancock (2007). Gender Classification using Shape from Shading, in *Proceedings of the British Machine Vision Conference (BMVC) 2007*. 29, 33, 95

P. Yan, S. Khan, and M. Shah (2007). 3D Model based Object Class Detection in An Arbitrary View, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. 25, 33, 109, 110, 113

B. Yao and L. Fei-Fei (2010). Grouplet: a Structured Image Representation for Recognizing Human and Object Interactions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. 147

T. Zesch and I. Gurevych (2010). Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words, *Journal of Natural Language Engineering*, vol. 16. 45, 129, 141

J. Zhang, K. Siddiqi, D. Macrini, A. Shokoufandeh, and S. Dickinson (2005). Retrieving Articulated 3-D Models Using Medial Surfaces and Their Graph Spectra, in *Energy Minimization Methods in Computer Vision and Pattern Recognition, 5th International Workshop 2005*. 142

Q. Zhu, G. Song, and J. Shi (2007). Untangling Cycles for Contour Grouping, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. 19

Q. Zhu, L. Wang, Y. Wu, and J. Shi (2008). Contour Context Selection for Object Detection: A Set-to-Set Contour Matching Approach, in *Proceedings of the European Conference on Computer Vision (ECCV) 2008*. 18, 19, 81

S. Zhu (1999). Embedding Gestalt Laws in Markov Random Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 20, 76

S. Zhu, R. Zhang, and Z. Tu (2000). Integrating Bottom-Up/Top-Down for Object Recognition by Data Driven Markov Chain Monte Carlo, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2000*. 30, 76, 79, 114

M. Zillich (2007). Incremental Indexing for Parameter-Free Perceptual Grouping, in *31st Workshop of the Austrian Association for Pattern Recognition 2007*. 68

A. Zweig and D. Weinshall (2007). Exploiting Object Hierarchy: Combining Models from Different Category Levels, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. 2, 3, 37, 38, 74, 76, 124

CURRICULUM VITAE

Michael Stark

Born in Mainz, Germany

Education:	2006–2010	**TU Darmstadt, Germany** PhD student at the Multimodal Interactive Systems Group of Prof. B. Schiele
	1999–2005	**TU Darmstadt, Germany** Studies of Computer Science, graduation with degree *Dipl.-Inform.* Minor subjects: Computer and Communication Systems Diploma thesis: Combining Streaming and Materialization for Processing XQuery, supervised by Dr. Peter Fankhauser, Mary F. Fernandez, Ph.D., and Prof. Thomas Hofmann
Experience:	2005–2006	Research Associate at Fraunhofer Integrated Publication and Information Systems Institute (IPSI), division Data Mining (MINE), supervised by Dr. Ulrike von Luxburg and Prof. Thomas Hofmann, Darmstadt, Germany
	2005	Research Internship at AT&T Labs Research, Florham Park, NJ, USA
	2003–2004	Student research assistant at Fraunhofer IPSI, division Open Adaptive Information Management Systems (OASYS), supervised by Dr. Peter Pankhauser, Darmstadt, Germany
	2002	Student research assistant at Fraunhofer IPSI, division Publication Engineering and Design (TOPAS), supervised by Prof. Klaus Mätzel, Darmstadt, Germany

PUBLICATIONS

[10] *"Combining Language Sources and Robust Semantic Relatedness for Attribute-Based Knowledge Transfer"*
Marcus Rohrbach, Michael Stark, György Szarvas, Bernt Schiele
In First International Workshop on Parts and Attributes (**PnA2010**), in conjunction with **ECCV** 2010, Crete, Greece, 2010

[9] *"Back to the Future: Learning Shape Models from 3D CAD Data"*
Michael Stark, Michael Goesele, Bernt Schiele
In British Machine Vision Conference (**BMVC**), Aberystwyth, UK, 2010

[8] *"What Helps Where – And Why? Semantic Relatedness for Knowledge Transfer"*
Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, Bernt Schiele
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), San Francisco, CA, 2010

[7] *"Categorical Perception"*
Mario Fritz, Mykhaylo Andriluka, Sanja Fidler, Michael Stark, Ales Leonardis, Bernt Schiele
In Categorical Perception, Cognitive Systems Monographs (8), Springer, 2010

[6] *"Multi-Modal Learning"*
Danijel Skocaj, Matej Kristan, Alen Vrecko, Ales Leonardis, Mario Fritz, Michael Stark, Bernt Schiele, Somboon Hongeng, Jeremy L. Wyatt
In Categorical Perception, Cognitive Systems Monographs (8), Springer, 2010

[5] *"Shading Cues for Object Class Detection"*
Michael Stark, Michael Goesele, Bernt Schiele
In International IEEE Workshop on 3D Representation for Recognition (**3dRR-09**), in conjunction with **ICCV** 2009, Kyoto, Japan, 2009

[4] *"A Shape-Based Object Class Model for Knowledge Transfer"*
Michael Stark, Michael Goesele, Bernt Schiele
In IEEE International Conference on Computer Vision (**ICCV**), Kyoto, Japan, 2009

[3] *"Functional Object Class Detection Based on Learned Affordance Cues"*
Michael Stark, Philipp Lies, Michael Zillich, Jeremy Wyatt, Bernt Schiele
In International Conference on Computer Vision Systems (**ICVS**), Santorini, Greece, 2008

[2] *"How Good are Local Features for Classes of Geometric Objects"*
Michael Stark, Bernt Schiele
In IEEE International Conference on Computer Vision (**ICCV**), Rio de Janeiro, Brazil, 2007

[1] *"XQuery Streaming à la Carte"*
Mary F. Fernández, Philippe Michiels, Jérôme Siméon, Michael Stark
In IEEE International Conference on Data Engineering (**ICDE**), Istanbul, Turkey, 2007